FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Gestão de Descarregamentos: Um estudo empírico de várias abordagens baseadas em regressão

Susana Oliveira Pacheco Neves

Mestrado Integrado em Engenharia Eletrotécnica e de Computadores

Orientador: Prof. Carlos Manuel Milheiro de Oliveira Pinto Soares

Co-orientador: Engº Nuno Guedes

28 de Julho de 2015

U.PORTO
FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO

MIEEC - MESTRADO INTEGRADO EM ENGENHARIA ELETROTÉCNICA E DE COMPUTADORES | **2014/2015**

A Dissertação intitulada

"Gestão de Descarregamentos: Um estudo empírico de várias abordagens baseadas em regressão"

foi aprovada em provas realizadas em 21-07-2015

o júri

Presidente Professor Doutor António José de Pina Martins
Professor Auxiliar do Departamento de Engenharia Eletrotécnica e de Computadores da Faculdade de Engenharia da Universidade do Porto

Professora Doutora Rita Paula Almeida Ribeiro
Professora Auxiliar Convidada do Departamento de Ciência de Computadores da Faculdade de Ciências da Universidade do Porto

Professor Doutor Carlos Manuel Milheiro de Oliveira Pinto Soares
Professor Associado do Departamento de Engenharia Informática da Faculdade de Engenharia da Universidade do Porto

O autor declara que a presente dissertação (ou relatório de projeto) é da sua exclusiva autoria e foi escrita sem qualquer apoio externo não explicitamente autorizado. Os resultados, ideias, parágrafos, ou outros extratos tomados de ou inspirados em trabalhos de outros autores, e demais referências bibliográficas usadas, são corretamente citados.

Susana Neves

Autor - Susana Oliveira Pacheco Neves

Faculdade de Engenharia da Universidade do Porto

Resumo

Os aproveitamentos hidroelétricos têm vindo a ocupar uma posição de particular importância na produção de energia em Portugal e noutros países. Devido ao seu crescimento em número e relevância, as barragens têm levantado vários problemas relacionados com a sua construção e operação, sendo um destes a gestão dos descarregamentos. Os descarregamentos consistem na libertação de água através da barragem e são indispensáveis para o correto funcionamento desta. Uma boa gestão dos descarregamentos proporciona uma série de benefícios, não só para a empresa responsável mas também para as populações vizinhas e para o ambiente. No entanto, em épocas de cheia, esta tarefa torna-se particularmente difícil, devido aos elevados afluentes, sendo benéfico a existência de ferramentas auxiliares na decisão dos descarregamentos.

De forma a dar resposta a este problema, seguiram-se algumas abordagens pertinentes com objetivo de prever o estado futuro das variáveis hidrológicas importantes na gestão de descarregamentos, aplicando técnicas de *Data Mining* a dados relativos a um par de barragens em cascata. Mais concretamente, utilizou-se dados horários e alguns dados esporádicos de duas barragens localizadas consecutivamente, e recorreu-se à análise de séries temporais, regressão de quantis e previsão de valores raros extremos, de forma a desenvolver modelos que permitissem a previsão do caudal afluente ao longo de um horizonte de previsão de 10 horas. Estas duas últimas abordagens são relevantes pois, dado que o foco está nas épocas de cheias e não nas situações consideradas normais, é importante o estudo de métodos próprios de casos raros extremos. É também de salientar que, segundo a pesquisa elaborada, estes dois métodos não foram ainda aplicados à previsão de descarregamentos, sendo por isso, abordagens mais inovadoras e desafiantes comparativamente às outras. Desenvolveu-se também uma nova abordagem, tirando partido do estudo feito anteriormente mas com algumas alterações relativamente ao conceito, fazendo com que esta passasse a ser uma abordagem de simulação.

Apesar de não se ter alcançado o desempenho pretendido em cheia, foi possível retirar diversas conclusões interessantes e até mesmo obter resultados aceitáveis no caso geral. Nomeadamente, a última abordagem referida permitiu os obter melhores resultados em cheia, comparativamente com as anteriores, sendo estas previsões melhores que as do modelo *naive*.

Abstract

Hydroelectric plants have come to occupy a position of particular importance in energy production in Portugal and other countries. Due to the growth in number and relevance, dams have raised several issues related to their construction and operation, one of them being the management of releases. Releases are indispensable for the correct operation of a dam and their good management provides a number of benefits, not only for the company responsible but also for the local communities and the environment. However, in the rainy season and with high tributary flows, this task becomes particularly difficult, and the existence of auxiliary tools for releases management is benefic.

In order to address this problem, several approaches were developed to predict the future state of important hydrological variables in the release decisions. To create these models, several Data Mining techniques were applied on data from a pair of reservoirs in cascade. Specifically, hourly and sporadic data were used, from two reservoirs located consecutively, and the models were developed using time series analysis, quantile regression and prediction of rare extreme values, allowing the forecast of the tributary flow over a 10 hour prediction horizon. These last two approaches are relevant because, since the focus is on floods seasons and not in normal situations, it is important to study proper methods for extreme and rare cases. It should also be noted that, according to the search elaborated, these two methods have not been implemented on the release prediction yet, being therefore more challenging and innovative approaches than the others. A new approach was also developed, taking advantage of the previous study but with some changes in the concept, causing this to be a simulation approach.

Although the desired performance was not achieved in the flood cases, several interesting conclusions can be drawn and even acceptable results are presented in the general case. The last mentioned approach made possible to obtain better results in the flood cases, comparing with the previous ones, and better predictions than the ones in the naive model.

Agradecimentos

A realização deste trabalho só foi possível graças a todas as pessoas e entidades que, direta ou indiretamente, contribuíram para o meu percurso pessoal e académico, tendo-me ajudado a superar este desafio. Às seguintes pessoas gostaria de deixar um agradecimento especial:

Ao meu orientador Prof. Carlos Manuel Milheiro de Oliveira Pinto Soares pelo seu indispensável apoio, pela simpatia demonstrada, pelo conhecimento transmitido e pelo incentivo prestado, sobretudo nas alturas de maior frustração.

Ao meu co-orientador Engº Nuno Guedes pela sua constante disponibilidade quer para a transmissão de informações quer para a passagem do material necessário à realização desta tese.

Ao Engº Pacheco de Andrade e à EDP - Gestão da Produção de Energia pela oportunidade e privilégio de realizar a tese com a parceria de uma empresa Portuguesa de referência.

Aos meus amigos, que me acompanharam ao longo deste percurso e que me irão acompanhar nos próximos, pela sua inestimável amizade e pelo seu apoio nos bons e maus momentos.

Ao meu namorado, pelo seu carinho, pela paciência e encorajamento que me fizeram continuar sem desanimar. E pelas suas piadas que me alegraram até nos momentos mais sombrios.

Por fim, apesar de não haver palavras que possam alguma vez exprimir o que fizeram por mim, a toda à minha família pelo seu amor e aprovação incondicionais e por me terem guiado até este momento, e pelos mais que hão de vir. Ao meu pai por zelar sempre por mim, tanto nas suas ações como nas suas palavras (mesmo que às vezes pessimistas), e pela sua ajuda e apoio sem as quais não me teria sido possível alcançar os meus objetivos. À minha mãe pelo seu carinho, pela sua amizade e por todas as suas ações, grandes e pequenas, que a tornam numa pessoa indispensável para a minha vida e sem a qual não seria quem sou.

Susana Neves

"An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem."

John Tukey

Conteúdo

Lista de Figuras

Lista de Tabelas

Abreviaturas e Símbolos

Abreviaturas

APA	Agência Portuguesa do Ambiente
AR	*Autoregressive*
CRF	*Conditional inference Random Forests*
CRISP-DM	*CRoss-Industry Standard Process for Data Mining*
DYNLM	*Dynamic Linear Model*
EDP	Energias de Portugal
MA	*Moving Average*
MAPE	*Mean Absolute Percentage Error*
MSBVAR	*Markov-Switching, Bayesian, Vector Autoregression Models*
MTS	*Multivariate Time Series*
NN	*Neural Network*
RAN	Regulador Automático de Nível
RF	*Random Forests*
RMSE	*Root Mean Squared Error*
RMSLE	*Root Mean Squared Logarithmic Error*
SMOTE	*Synthetic Minority Over-sampling Technique*
SZBVAR	*Reduced form Sims-Zha Bayesian VAR*
VAR	*Vector Autoregressive*
VARMA	*Vector Autoregressive Moving-Average*
VMA	*Vector Moving Average*

Símbolos

Q	Caudal
V	Volume

Capítulo 1

Introdução

Neste capítulo faz-se uma breve introdução aos aproveitamentos hidroelétricos em Portugal, relativamente à sua história, importância e proveitos. É também apresentado o enquadramento da dissertação neste panorama e os seus objetivos. Por último, na secção 1.3, é feita uma descrição da estrutura do documento, salientando os pontos mais importantes de cada capítulo.

1.1 Aproveitamentos Hidroelétricos em Portugal

A água sempre foi considerada um bem essencial à vida humana e utilizada nas mais diversas áreas. Nos últimos séculos, acrescentou-se um novo papel a este recurso: a produção de energia elétrica através da força motriz da água. Em Portugal, a produção de energia elétrica através da água dos rios teve início na última década do século XIX. No entanto, os grandes aproveitamentos hidroelétricos só começaram a surgir em 1950, tendo sofrido um crescimento significativo desde então [1]. Nos dias de hoje, a produção hidroelétrica toma um papel fundamental na produção de energia elétrica em Portugal pois, para além de ser um recurso energético renovável, não poluente e endógeno, contribui com cerca de 30% do consumo energético mensal [2].

Para além do seu peso no consumo energético, a grande hídrica é a principal responsável por cobrir a ponta do diagrama de carga, ou seja, garante a produção de energia elétrica suficiente nas horas de maior consumo. Isto deve-se ao facto de, ao contrário de outras energias renováveis, como a energia eólica, que não podem ser controladas, as barragens permitem a acumulação de água na albufeira, armazenando-se assim energia potencial gravítica, que pode ser transformada em energia elétrica quando necessário.

A capacidade de armazenamento das albufeiras permite não só compensar estas variações diárias como também assegurar o abastecimento em situações ocasionais, contribuindo para a gestão de outras fontes de energia renovável mais variáveis. Por exemplo, quando a eólica está a produzir significativamente mas o consumo de energia pela população é baixo as centrais hidroelétricas podem consumir essa energia extra para bombar água para montante. Quando a eólica produz pouca energia, a hídrica pode produzir mais para compensar, complementando-se assim uma à outra. Mas as vantagens de um aproveitamento hidroelétrico não se resumem apenas à produção

de energia pois este é também capaz de abastecer água às populações e indústrias vizinhas, contribuir para minimizar os efeitos nocivos de situações de cheias e secas e facilitar a navegabilidade comercial, turística e as práticas de lazer [1]. Em suma, os aproveitamento hidroelétricos são não só uma fonte de energia renovável controlável como também uma mais valia para as comunidades circundantes à albufeira.

1.2 Motivação e objetivos

O contexto geográfico, ecológico e económico-social pode variar significativamente entre aproveitamentos hidroelétricos. Para além disso, a estrutura da barragem é determinante no modo como esta é explorada. De uma maneira muito genérica, pode-se distinguir os aproveitamentos hidroelétricos em fio de água ou albufeira, mediante a sua capacidade de retenção de água. O primeiro caso dá-se em locais cujo declive é pouco acentuado, não permitindo grandes armazenamentos e obrigando a que quase toda a água que chega a montante seja imediatamente lançada para jusante. No caso de aproveitamentos com albufeira, existe uma retenção de água no reservatório a montante da barragem, permitindo uma melhor gestão da produção de energia e regularização do regime dos rios. Nestas barragens, havendo retenção significativa de água, existe também uma preocupação com os descarregamentos, que consistem na libertação de água que, não sendo utilizada para produzir energia, permite controlar o caudal e cota do rio, tanto a montante como a jusante. Assim, a sua gestão é realmente relevante nos aproveitamentos com albufeira [3].

Para além disso, os descarregamentos tomam principal importância nas épocas de cheias e secas, pois permitem uma regularização do caudal e cotas do rio, funcionando como um mecanismo de defesa contra variações extremas. Em época de chuva intensa, o caudal libertado deve ser suficiente para que a cota máxima a montante não seja excedida, o que poria em risco a segurança da barragem. Simultaneamente, este não pode ser tão elevado que provoque inundações a jusante. Na época seca, deve-se tentar acumular o máximo de água que for possível, libertando pelo menos o suficiente para cumprir com os requisitos ecológicos. Todas estas decisões são baseadas em normas de descarregamentos, impostas a cada barragem, que permitem uma regulação sistemática e controlada [4].

Atualmente, é apenas possível obter informação sobre o estado presente da própria barragem e das barragens vizinhas, permitindo uma abordagem reativa face a esta informação. As normas de descarregamento [5, 6, 7, 8, 9] utilizam apenas a informação relativa à própria barragem e devem garantir sempre a segurança da mesma. No entanto, em épocas de cheias o comportamento do rio pode tornar-se imprevisível e rapidamente variável, mesmo com o auxílio da barragem. Nestas alturas, o recurso a outras ferramentas, para além do normativo existente, que auxiliem a decisão dos descarregamentos revela-se fundamental, possibilitando uma gestão mais eficiente. A informação do estado atual de barragens vizinhas, aliada à experiência e sensibilidade do operador, pode permitir a tomada de ações pro-ativas. Contudo, esta abordagem requer algum conhecimento sobre o tempo que leva entre um descarregamento e a chegada desse caudal às barragens a jusante.

Dado que esta informação é pouco precisa e nem sempre se encontra disponível, deve-se procurar encontrar outros métodos que auxiliem a Gestão de Descarregamentos.

Procurando dar resposta a este problema, esta dissertação tem como objetivo auxiliar as decisões de descarregamentos, baseando-se no estudo de comportamentos passados. A sua elaboração conta com a colaboração da EDP - Gestão da Produção de Energia, que forneceu fornecer os dados e informações necessárias ao seu desenvolvimento.

Para alcançar o objetivo proposto é necessário conjugar duas áreas de estudo: a Gestão de Reservatórios de água e o *Data Mining*. A Gestão de Reservatórios de água pretende encontrar formas de controlar a retenção e libertação de água nestes reservatórios, podendo ter em vista critérios como a optimização da produção [10], a preservação dos ecossistemas [11], a irrigação de plantações [12], etc. Por sua vez, o *Data Mining* é um processo que, a partir de grandes quantidades de dados, procura extrair padrões e/ou relações sistemáticas entre variáveis, aplicando posteriormente este conhecimento a novos dados [13]. O *Data Mining* pode ser aplicado à Gestão de Reservatórios de água na medida em que permite não só prever mas também compreender melhor as causas e consequências de alguns fenómenos.

O objetivo é então a aplicação de técnicas de *Data Mining*, a dados hidrológicos relativos às barragens de Venda Nova e Alto Rabagão, de forma a construir um modelo capaz de prever o estado de uma ou mais variáveis importantes na Gestão de Descarregamentos, em situações de cheias. Redefinindo este objetivo em termos académicos, pode-se dizer que esta tese tem em vista o estudo, implementação e comparação de diferentes métodos de *Data Mining*, nomeadamente séries temporais, regressão de quantis e previsão de valores raros extremos.

1.3 Estrutura da Dissertação

O restante relatório desta dissertação encontra-se divido em mais cinco capítulos. O Capítulo 2 engloba toda a revisão bibliográfica, dividindo-se em três grandes secções, focando os métodos atuais de Gestão de Descarregamentos, uma pequena introdução ao processo de *Data Mining* e, por último, uma revisão dos trabalhos existentes na área, juntamente com as abordagens propostas. No Capítulo 3, faz-se uma análise aos dados disponíveis, salientando as suas características mais relevantes e algumas conclusões úteis para o desenvolvimento do trabalho. No Capítulo 4 apresenta-se todos os resultados obtidos com a abordagem de previsão, sendo de salientar as secções de séries temporais, regressão de quantis e previsão de valores raros extremos. O Capítulo 5 é semelhante ao Capítulo 4 mas relativo à abordagem de simulação. Por fim, no Capítulo 6 reúnem-se todas as conclusões parciais, fazendo uma apreciação global ao trabalho realizado e apresentando as sugestões de trabalho futuro.

Capítulo 2

Revisão Bibliográfica

Este capítulo engloba toda a revisão bibliográfica, constituindo um estudo das temáticas mais relevantes para a Gestão de Descarregamentos e o *Data Mining*, dividindo-se em três grandes secções. Na secção 2.1, são abordadas algumas questões relativas à Gestão de Descarregamentos, nomeadamente os seus objetivos e normas implementadas. É feito um estudo mais detalhado sobre a Gestão de Descarregamentos de Venda Nova, que servirá de caso de estudo para este trabalho. Na secção 2.2 é feita uma breve introdução ao *Data Mining* e à sua metodologia. Em seguida, na secção 2.3, é apresentada uma síntese de alguns trabalhos e abordagens existentes sobre a Gestão de Descarregamentos com *Data Mining*, fazendo uma divisão pelas várias categorias a distinguir. São também apresentadas as abordagens propostas. Por fim, na secção 2.4 é apresentado um resumo das conclusões mais importantes e é feita a ponte entre os métodos estudados previamente e a abordagem que este trabalho pretende tomar.

2.1 Gestão de descarregamentos

2.1.1 Objetivos

Como referido na secção 1.2, os descarregamentos fazem parte do quotidiano de uma barragem e são indispensáveis para o seu correto funcionamento. A gestão destes não se trata de uma tarefa trivial, pois é necessário criar um equilíbrio entre a necessidade de reter água para posterior produção de energia e as questões de segurança e correto aproveitamento do rio.

Para que os descarregamentos sejam efetuados corretamente, existem normas, validadas pela APA (Agência Portuguesa do Ambiente), que indicam como estes devem ser realizados [6, 7, 9]. Estas normas são relativas a cada barragem e têm em vista:

- Garantir que a cota máxima não é ultrapassada, mesmo em alturas de cheia.
- Permitir que os descarregamentos sejam feitos de forma gradual e que o caudal descarregado seja o mais bem distribuído possível pelas suas comportas.
- Minorar os efeitos nocivos provocados pelo caudal do rio, como por exemplo, a erosão das margens.

- Proteger as populações a jusante e minimizar a frequência e gravidade de eventuais cheias a montante.

- Possibilitar a navegabilidade em segurança, a jusante da barragem.

As normas tornam possível assegurar todos estes benefícios e outros que sejam específicos de cada barragem. O não cumprimento destas normas pode levar a consequências desastrosas, pondo mesmo em risco a segurança da barragem e das populações vizinhas.

2.1.2 Conceitos importantes

Antes de passar à descrição das normas e gestão dos descarregamentos é necessário cobrir alguns conceitos técnicos fundamentais. Em primeiro lugar, é preciso fazer uma breve explicação do funcionamento de uma central hidroelétrica, como a esquematizada na figura 2.1.

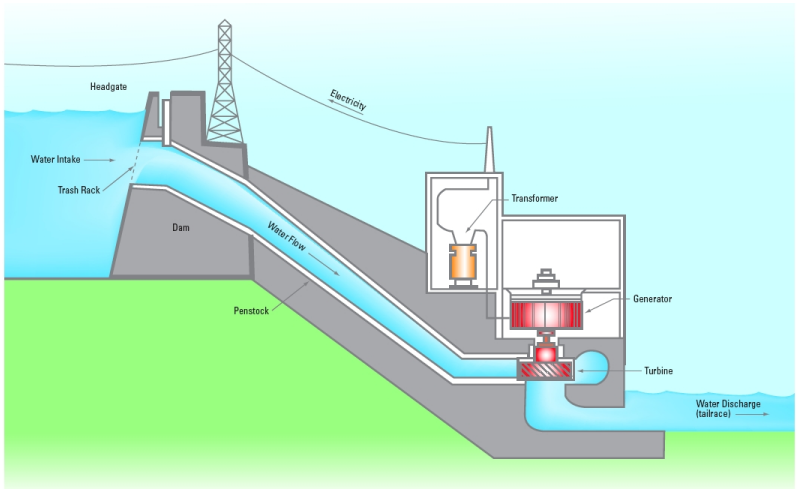


Figura 2.1: Esquema transversal de uma central hidroelétrica [14]

A diferença entre o nível de água a montante e o nível a jusante origina uma energia potencial gravítica, que é tanto maior quanto maior for este desnível. A energia potencial faz com que a água acumulada a montante flua a elevadas velocidades quando libertada. Este é o fenómeno que permite a produção de energia hidroelétrica. A água, quando conduzida pelo circuito hidráulico até à turbina, faz girar as pás, devido à força cinética que lhe é impressa. Por sua vez, a turbina está ligada a um alternador que permite converter a energia mecânica, produzida pelo movimento das pás, em energia elétrica. Este é o princípio base de produção de energia numa barragem.

Quanto mais água estiver acumulada a montante, mais energia se poderá produzir. No entanto, existem regras para o que se pode armazenar e o que se deve descarregar, sendo a gestão de água na albufeira uma questão regulada, mas complexa. Assim, e como foi referido anteriormente, os descarregamentos são essenciais no do dia-a-dia das barragens e a sua boa gestão permite um maior aproveitamento dos recursos do rio, tornando a central mais rentável.

Para se poder estudar a Gestão de Descarregamentos é necessário primeiro conhecer as variáveis associadas à situação hidrológica de uma barragem:

- Cota Montante/Jusante - Nível da água do rio (m) a montante/jusante da central.
- Caudal Afluente - Quantidade de água (m^3/s) que chega a montante da barragem.
- Caudal Turbinado - Quantidade de água (m^3/s) que é passada de montante para jusante no processo de produção de energia.
- Caudal Descarregado - Quantidade de água (m^3/s) que é passada de montante para jusante pelos descarregadores.
- Caudal Ecológico - Quantidade de água (m^3/s) que é passada de montante para jusante por um canal próprio que assegura o caudal mínimo a libertar.
- Caudal Bombado - Quantidade de água (m^3/s) que é passada de jusante para montante durante a bombagem. Este processo utiliza a energia em excesso na rede de forma a acumular mais água na albufeira para posterior produção de energia.

Nem todas as barragens possuem bombagem ou um canal próprio para o caudal ecológico, como por exemplo as de fio de água, sendo portanto o cálculo do caudal (Q) afluente diferente conforme o caso em questão. Se considerarmos duas barragens, uma primeira com bombagem (1) e a imediatamente a jusante (2) desta, o caudal afluente nestas barragens é dado por:

$$Q_{Afluente_1} = Q_{Turbinado_1} + Q_{Descarregado_1} \pm \frac{\Delta V_1}{\Delta t_1} \tag{2.1}$$

$$Q_{Afluente_2} = Q_{Turbinado_2} + Q_{Descarregado_2} + Q_{Bombado_1} \pm \frac{\Delta V_2}{\Delta t_2} \tag{2.2}$$

A variação de volume (V) traduz o aumento ou diminuição do volume de água presente na albufeira, e está diretamente relacionado com a alteração do caudal afluente. Intuitivamente, se todas as outras variáveis se mantiverem constantes, um aumento no caudal afluente traduz um aumento no volume de água da albufeira.

Nem todas as variáveis são da mesma natureza, umas são mais imprevisíveis que outras. Em primeiro lugar, é preciso distinguir que existem algumas variáveis que são controladas pelos operadores, sendo estas o caudal turbinado, descarregado, bombado e ecológico. As restantes variáveis não são diretamente controláveis mas influenciam as decisões de descarregamento e são influenciadas por estas. Apesar de estas variáveis estarem todas relacionadas umas com as outras, sobretudo em épocas de cheia pode dizer-se que as variáveis não controláveis são mais imprevisíveis, dado as restantes são da responsabilidade do operador.

Para todas as barragens é feito o registo horário destas variáveis, constituindo assim uma base de dados histórica, útil não só para monitorização e controlo mas também para estudos futuros, como este.

2.1.3 Normas

Em primeiro lugar, é preciso compreender que cada aproveitamento hidrológico tem as suas próprias características e condicionantes geológicas, hidrológicas, sociais e estruturais. Assim, torna-se necessário que cada barragem tenha as suas próprias normas de descarregamento, adaptadas para a situação em questão. Portanto, optou-se por primeiro abordar este tópico de uma forma mais global, não entrando em pormenores que seriam específicos de cada barragem, e na subsecção 2.1.3.1 aprofundar esta questão apenas para uma das barragens envolvidas a nível prático.

De uma forma geral, para caudais afluentes inferiores a um dado limite, os descarregamentos são efectuados de forma automática pelo Regulador Automático de Nível (RAN). Este sistema está incorporado num autómato e, conforme o nível de referência pretendido e o estado dos outros parâmetros hidrológicos que influenciam os descarregamentos, comanda a abertura das várias comportas, seguindo a sequência de abertura estipulada [6, 7, 9]. Quando o caudal afluente é superior ao limite estabelecido para o uso do RAN, o comando das comportas passa a ser feito de forma manual. O operador baseia-se em documentos com as normas de descarregamento impostas para aquela barragem, como por exemplo, as sequências de manobras, que indica qual a abertura de cada descarregador, conforme a situação.

Em seguida é dado um exemplo específico das normas de descarregamento da barragem de Venda Nova, para a qual se pretende fazer as previsões.

2.1.3.1 Normas de Descarregamento em Venda Nova

Venda Nova é um aproveitamento hídrico com albufeira [8] no rio Rabagão, localizada em Vila Nova e pertencente ao centro de produção do Cávado-Lima. A sua entrada em serviço deu-se em 1951 e produz em média 439 GWh por ano. Tal como se pode observar na figura 2.2, Venda Nova possui um descarregador de cheias, com duas comportas, e um de fundo, com uma comporta. Estes são capazes de descarregar, no máximo, $1100 m^3/s$ e $130 m^3/s$, respectivamente, sendo os grandes descarregamentos assegurados pelo descarregador de cheias. O descarregador de fundo é utilizado para esvaziar a albufeira mas, por causa do tipo de construção, não deve funcionar em simultâneo com o descarregador de cheias.

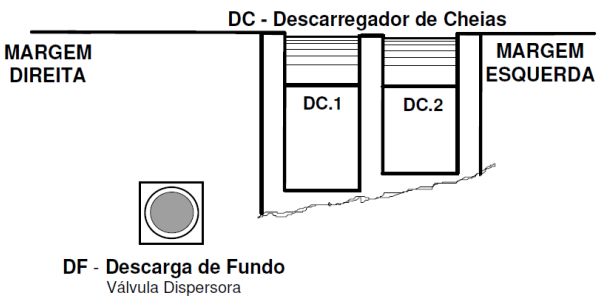


Figura 2.2: Órgãos de descarga de Venda Nova [8]

As normas de descarregamento desta barragem prevêem que seja assegurada, normalmente, uma cota montante de $699m$ no Inverno (01/10 a 31/3) e 700 m no verão (1/6 a 30/7), podendo variar linearmente entre estes níveis no restante tempo. O nível máximo de cheia está fixado nos $700,5m$. Os operadores devem atuar de forma a cumprir com estes limites, prestando especial atenção às épocas de cheia pois, tal como foi salientado na secção 1.2, é quando esta gestão se revela mais preocupante. Para estas alturas, as ações definidas pelas normas de descarregamento variam conforme o nível atual da albufeira, contemplando as seguintes situações:

1. Nível da Albufeira menor a $699m$: Não é necessário descarregar.
2. Nível da Albufeira igual a $699m$: O caudal descarregado deve ser igual ao afluente, de forma a manter o nível da albufeira no valor atual.
3. Nível da Albufeira superior a $699m$: Deve-se proceder às sequências regularizadoras da albufeira.

O terceiro ponto é de maior importância pois, para além de ser o mais relevante para este trabalho dado corresponder à altura de tomar decisões de descarregamento, é também nesses casos que o comportamento das variáveis hidrológicas se torna mais imprevisível. Nesta situação pode-se chegar a ultrapassar o nível máximo de cheia estipulado, sendo isso indesejável como até mesmo potencialmente perigoso.

A tabela de sequência de manobras, que se pode consultar no Anexo A, permite ao operador saber qual será a próxima ação a tomar, dependendo da situação atual e do regime afluente. Em regime de afluente crescente, ou seja, quando o caudal afluente está a aumentar, deve-se executar a sequência de manobras apresentada, ou um subconjunto destas, pela sua ordem. Caso contrário, em regime de afluente decrescente, deve ser seguida a ordem inversa. Por exemplo, se atualmente as duas comportas dos descarregadores de cheia se encontram com uma abertura de um metro, correspondente à situação 4, deve-se passar à situação 5 em regime de afluente crescente, ou à situação 3 em regime decrescente. De notar que não é obrigatório que as manobras sejam feitas de forma estritamente sequencial. Pode-se saltar um ou mais passos, desde que não se atinjam variações superiores a $240m^3/s$ na fase crescente e $180m^3/s$ na fase decrescente.

Estas normas constituem o panorama geral das ações a tomar em Venda Nova, podendo-se dizer que constituem uma abordagem reativa face ao estado atual da própria barragem.

2.1.4 Previsão de descarregamentos

Algumas barragens possuem uma estimativa do tempo que decorre entre um descarregamento efetuado numa barragem mais a montante e a chegada desse caudal à barragem em questão [6]. Atualmente, esta informação é a única maneira de prever o caudal afluente e, a partir daí, tomar medidas preventivas que permitam minimizar as variações acentuadas no caudal descarregado. Este tipo de previsão é muito genérico e pouco adaptável, sendo suficiente para alguns casos pontuais mas falível em situações mais complexas, nomeadamente, em alturas de cheias. Para

além disso, não existe estimativa para todos os pares de barragem, nomeadamente, para o par de estudo de Venda Nova e Alto Rabagão.

Para além desta informação, o operador tem também acesso às curvas de regolfo de cada barragem [6, 7]. Estas curvas são uma representação gráfica que, para um conjunto de caudais afluentes, fornecem informação da cota do rio a montante da barragem, em função da distância à origem. Sabendo a cota em vários pontos a montante da barragem, o operador poderá intuitivamente prever o comportamento futuro do rio, e tomar decisões antecipando esse comportamento. Mais uma vez, este método é muito genérico, não cobrindo situações extremas. A curva de regolfo relativa à barragem do Pocinho pode ser consultada no Anexo A.2.

Como foi sendo salientado, as situações de grandes caudais afluentes são especialmente preocupantes, existindo uma necessidade de obter previsões melhoradas para estes casos. Assim, aproveitando esta oportunidade de estudo e desenvolvimento, tenciona-se implementar um ou mais modelos, baseados na aplicação de uma dada abordagem de *Data Mining* a dados históricos, que permitam averiguar se é possível prever o caudal afluente em situações de cheia com uma dada confiança e, em caso afirmativo, criar uma aplicação que suporte esse processo. Esta previsão irá então permitir aos operadores tomarem decisões melhoradas relativamente aos descarregamentos.

2.2 Introdução ao *Data Mining*

> "*Data mining, also popularly referred to as knowledge discovery from data (KDD), is the automated or convenient extraction of patterns representing knowledge implicitly stored or captured in large databases, data warehouses, the Web, other massive information repositories, or data streams.*" [15, chap. Preface]

Conforme descrito em [15], de uma forma simplista pode-se definir *Data Mining* como o processo de extração de conhecimento a partir de grandes quantidades de dados. Este processo pode ser dividido genericamente em duas categorias: descritivo ou preditivo. O *Data mining* descritivo preocupa-se em caracterizar as propriedades gerais dos dados, enquanto que o *Data Mining* preditivo procura fazer inferências sobre estes, de modo a obter previsões.

Como referido na secção 1.2, o objetivo desta tese é o desenvolvimento de um modelo capaz de prever o caudal afluente de uma dada barragem. Este problema enquadra-se no *Data Mining* preditivo, mais especificamente na previsão numérica ou regressão. Neste caso, o caudal afluente é considerado o atributo dependente, por ser a variável a prever e por se supor que o seu valor depende de outras, sendo por isso a saída do modelo [16]. Por sua vez, as variáveis utilizadas para prever o atributo dependente são consideradas variáveis independentes e são a entrada do modelo.

De entre as etapas do processo de *Data Mining* preditivo pode-se salientar a aprendizagem e a previsão. Na fase de aprendizagem os dados de treino são analisados por algoritmos de regressão, retornando um modelo capaz de analisar vários atributos independentes e atribuir um valor ao atributo dependente. Na fase de previsão, o modelo criado anteriormente é aplicado a dados de teste, com o intuito de prever o valor do atributo dependente desses dados. É então avaliada a performance do modelo e, caso seja satisfatória, este é aplicado a novos dados [15].

Seguidamente, apresenta-se uma breve descrição do processo que se irá implementar, seguindo a metodologia CRISP-DM (*CRoss-Industry Standard Process for Data Mining*).

2.2.1 Etapas do Processo de *Data Mining*

O CRISP-DM [17], é uma metodologia *standard* para o desenvolvimento de projetos de *Data Mining*, criada em 1999 por um consórcio de empresas e profissionais nesta área. Segundo esta metodologia, o ciclo de vida de um projeto de *Data Mining* consiste essencialmente em seis fases não estanques, como se pode observar na figura 2.3, tendo por base o seguinte:

- Compreensão do negócio (*Business Understanding*): Adotar a perspetiva do negócio, por forma a compreender melhor o que o cliente pretende e quais os objetivos do trabalho.
- Compreensão dos dados (*Data Understanding*): Recolher, explorar e estudar os dados disponíveis e relevantes para o trabalho, de forma a familiarizar-se com estes e identificar eventuais problemas de qualidade.
- Preparação dos dados (*Data Preparation*): Série de tarefas que transforma os dados iniciais num conjunto final, preparado para ser posteriormente utilizado pelas ferramentas de modelação. Esta etapa é geralmente morosa e com grande impacto no resultado final.
- Modelação (*Modeling*): Aplicação de um ou vários algoritmos ao conjunto de dados desenvolvido anteriormente, calibrando os parâmetros e maximizando uma dada medida de

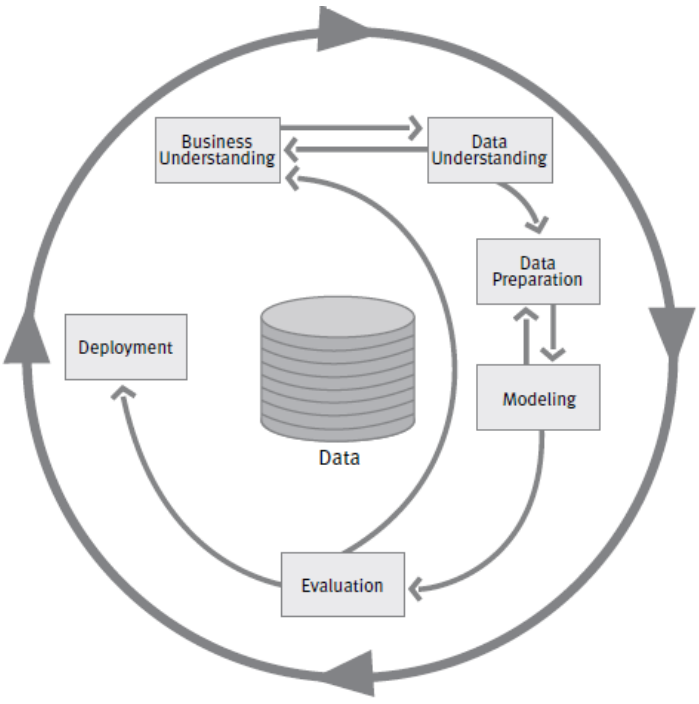


Figura 2.3: Fases de um projeto de *Data Mining* segundo o CRISP-DM [17]

desempenho. Nesta fase não é necessário investir grandes quantidades de tempo, se se souber quais os modelos adequados ao problema e como estes funcionam.

- Avaliação (*Evaluation*): Fase em que se avalia o desempenho dos modelos obtidos, face a um novo conjunto de dados, e em que se determina o grau de satisfação perante os objetivos de negócio estabelecidos inicialmente. Deve também ser feita uma análise crítica do trabalho realizado e planear quais os próximos passos.

- Implementação (*Deployment*): Transformação do modelo ou conhecimento obtidos em algo que o cliente possa compreender e integrar no seu sistema. Para além disso, deve-se pensar na melhor estratégia de monitorização e manutenção do resultado do projeto.

O cumprimento desta metodologia permite que o trabalho flua de uma forma estruturada, sabendo sempre quais são os objetivos atuais e o próximo passo. Isto não só permite a obtenção de melhores resultados como promove uma melhor gestão do tempo. Sendo assim, decidiu-se seguir estes passos, fazendo pequenas adaptações de forma a enquadrar melhor no trabalho em questão, estando estas fases integradas ao longo deste relatório. Em primeiro lugar, elaborou-se a compreensão do negócio, apresentada na secção 1.2. A compreensão dos dados será apresentada em detalhe no Capítulo 3, incluindo na secção 3.4 a limpeza dos dados, uma das tarefas da preparação dos dados. A restante preparação dos dados, modelação e avaliação serão apresentadas ao longo dos Capítulos 4 e 5. Relativamente à implementação, este tópico é discutido no Capítulo 6.

2.3 A Gestão de Descarregamentos e o *Data Mining*

Como descrito na subsecção 2.1.3, as regras de operação dos descarregamentos são, na sua grande maioria, textos, gráficos e tabelas que permitem ao operador determinar facilmente a ação que deve tomar em cada situação. Ainda assim, existe uma possibilidade de melhorar estas decisões, contribuindo para um aumento da estabilidade do caudal do rio e trazendo benefícios a nível económico, ambiental e social.

Consequentemente, a Gestão de Descarregamentos e o *Data Mining* são áreas com potencial para se conjugarem, existindo alguns estudos recentes que propõem novos métodos e modelos capazes de auxiliar e melhorar estas decisões. Visto os aproveitamentos hidroelétricos manterem registos de várias variáveis que influenciam as decisões de descarregamentos, faz sentido analisar estes dados de forma a aperfeiçoar ou até mesmo substituir os modelos teóricos atuais.

2.3.1 Métodos e Abordagens Estudadas

Por forma a averiguar a relação existente entre a gestão de descarregamentos e o *Data Mining*, analisou-se diversos trabalhos realizados na área, obtendo-se um panorama geral dos métodos já implementados e dos resultados obtidos. A metodologia seguida baseou-se numa análise comparativa dos vários trabalhos, de modo a categorizá-los de acordo com as características identificadas. Do que foi observado, pode-se afirmar que existem seis categorias a distinguir:

- Tarefa de *Data Mining*: Classificação, previsão numérica ou outra;
- Algoritmos utilizados: Por exemplo, árvores de decisão;
- Variável dependente: Por exemplo, a quantidade de água a libertar ou o número de comportas abertas;
- Variáveis independentes: Por exemplo, nível de água, afluente, descargas passadas, etc.;
- Granularidade dos dados: Diários, semanais ou anuais;
- Medida de avaliação do desempenho: Por exemplo, a raiz do erro quadrático médio.

De acordo com estas categorias, apresenta-se seguidamente os estudos analisados.

2.3.1.1 Tarefa de *Data Mining*

Uma primeira distinção a fazer entre os trabalhos estudados é a tarefa de *data mining* a que estes se propõem. Alguns estudos abordam o problema da Gestão de Descarregamentos como sendo um problema de previsão numérica [18, 19, 20], enquanto que outros optam por tratá-lo como sendo de classificação [18, 21, 22]. Existem ainda trabalhos que aplicam técnicas de *Data Mining* à Gestão de Descarregamentos, não com o intuito de fazer previsões mas sim de analisar o fenómeno e extrair informações úteis, como por exemplo quais as variáveis que mais influenciam as decisões de descarregamento [23]. Estes últimos tentam compreender melhor o funcionamento dos descarregamentos e a sua dependência com as outras variáveis hidrológicas, tendo concluído que as variáveis mais relevantes para um descarregamento num dado momento são os descarregamentos efectuados anteriormente e caudal afluente presente e passado. Este conhecimento não só é importante como vai de encontro ao que era esperado intuitivamente. A identificação da tarefa de *data mining* é o primeiro passo na delineação da abordagem do trabalho.

2.3.1.2 Algoritmos utilizados

Os algoritmos devem ser escolhidos em função da tarefa de *data mining* a realizar e de outros critérios específicos de cada trabalho. Em estudos cujo primeiro objetivo é a previsão dos descarregamentos, pode-se encontrar algoritmos de previsão, como as árvores de decisão [18, 22], a regressão linear [18], redes neuronais [21, 22], classificador de *Naive Bayes* [22] e *Support vector machines* [20]. Por sua vez, em trabalhos cujo intuito principal é compreender melhor as decisões de descarregamentos, implementam-se técnicas de *Feature importance* e *Feature selection*, como por exemplo *Mutual information value* [23] e *Maximum relevance minimum redundancy* [19], respectivamente. Estas técnicas permitem averiguar a quantidade de informação que uma variável independente contém sobre a variável dependente, determinando quais as variáveis mais importantes para o modelo, e possibilitam a criação de subconjuntos de dados com estas variáveis. Esta informação, para além de ser relevante para os trabalhos em questão, proporciona aos estudos futuros a aquisição de alguma sensibilidade relativamente às variáveis e ao fenómeno em si.

2.3.1.3 Variável dependente

Como referido na secção 2.2, a variável dependente é aquela que se pretende prever, ou seja, a saída do modelo. De uma forma geral, os estudos de Gestão de Descarregamentos que fazem previsão numérica usam como variável dependente a quantidade de água a libertar [18, 21, 22]. Por outro lado, os que optam por classificação apresentam maior diversidade, podendo alternar entre a quantidade de água a libertar, tratada de uma forma discreta, [22], isto é, dividindo a gama de valores possíveis num número finito de patamares, o número de comportas abertas [21] ou uma medida discreta da qualidade da decisão, como por exemplo, patamares da percentagem da procura de água que foi correspondida [18]. Pode-se dizer que a variável dependente traduz o objetivo do trabalho.

2.3.1.4 Variáveis independentes

Para criar um modelo capaz de prever uma dada grandeza, identificada anteriormente como variável dependente, é necessário outras variáveis que sirvam de entrada a essa modelo, sendo estas as variáveis independentes. Nos estudos analisados, estas variam consoante os dados disponíveis e o propósito do trabalho mas, normalmente, utiliza-se o caudal afluente [18, 21, 19, 20, 23], a cota montante ou quantidade de água disponível no reservatório [21, 21, 19, 20, 23] e os descarregamentos passados [19, 20, 23]. Em alguns estudos é ainda utilizada a precipitação [22, 19, 23].

Principalmente em trabalhos que focam a previsão, existem algumas considerações e variantes que podem ser adicionadas de forma a tornar os modelos mais próximos da realidade. Por exemplo, quando existem barragens em cascata, a quantidade de água libertada numa dada barragem influencia diretamente a cota e o caudal das barragens imediatamente a montante e a jusante desta. De forma a ter em consideração este fenómeno, foram feitos esforços para criar modelos que incorporassem múltiplas barragens [18] na previsão dos descarregamentos num dado local, em vez de considerar apenas a barragem em questão. No entanto, o número de variáveis necessárias para estes modelos é muito superior, o que implica um custo computacional maior do que para uma única barragem. Um fenómeno tão ou mais importante que a influência de outras barragens nas decisões de descarregamento é a passagem do tempo. De uma forma geral, podemos dizer que o estado das variáveis hidrológicas hoje depende do estado destas mesmas variáveis ontem e anteontem, e as decisões de descarregamento não são isoladas no tempo. Para além disso, uma descarga a montante leva um determinado tempo a fazer-se sentir a jusante. Portanto, uma consideração muito importante a ter na elaboração de um modelo de previsão de descarregamentos, ou outra variável hidrológica associada a estes, é o tempo. Vários estudos trabalharam no sentido de incorporar nos seus modelos informação passada, de forma a que as previsões fossem feitas não só com o estado atual das variáveis mas também com o estado passado [20, 21, 22]. De uma forma geral, a técnica mais utilizadas para incluir informação temporal no modelo é criar novas variáveis que correspondam a valores anteriores na série temporal. Esta transformação pode ser feita de uma forma sistemática, utilizando uma *Sliding Window* [21], com um dado tamanho w, que em cada instante t adiciona as variáveis correspondentes a $t-1,...,t-w$.

Tanto os sistemas de múltiplos reservatórios como a análise temporal são importantes e conduzem a resultados melhores e mais próximos da realidade. A simples inclusão de variáveis relativas ao período imediatamente anterior já se traduz numa melhoria significativa na performance do modelo. No entanto, tanto um como outro implicam um aumento de complexidade do modelo e maior custo computacional. É preciso encontrar um compromisso entre a complexidade e a performance desejada.

2.3.1.5 Granularidade dos dados

A granularidade dos dados traduz a periodicidade com que estes foram recolhidos, sendo que as séries temporais são geralmente amostradas com um período fixo. Ao observar os estudos passados, constatou-se que a granularidade pouco varia, utilizando-se normalmente dados mensais [18, 19, 20, 22, 23], semanais [19, 23] ou diários [21], sendo que a grande maioria recorre a dados mensais. O facto de estes trabalhos não recorrerem a dados horários significa que o interesse ao nível da aplicação prática não o exigia, sendo apenas necessário retirar conclusões diariamente, semanalmente ou mensalmente. No entanto, a falta de estudos em escalas temporais menores pode-se dever à periodicidade de recolha dos dados não o permitir, ou seja, não existir registos com granularidade mais fina nas barragens de estudo. Ainda assim, de uma forma geral, estes estudos apresentam bons resultados, permitindo concluir que, pelo menos a nível mensal, é possível fazer uma boa previsão de qual deverá ser o caudal descarregado.

2.3.1.6 Medida de avaliação do desempenho

Após a criação de qualquer modelo, é necessário avaliar e comparar a sua qualidade. Par tal, utilizam-se medidas de performance, que tipicamente variam consoante se esteja a tratar de um problema de classificação ou previsão numérica, sendo mais comum a *accuracy* [21] e o RMSLE (*Root Mean Squared Logarithmic Error*) [22, 19, 20], respetivamente. O RMSLE permite obter uma medida de desempenho mais independente da grandeza da variável, dado fazer uma transformação logarítmica, podendo ser facilmente interpretada. Esta avaliação permite, para um dado conjunto de dados e de algoritmos, fazer uma comparação do desempenho entre os diferentes modelos resultantes. Para além disso, se as fases de modelação e avaliação forem bem conduzidas, os resultados obtidos num determinado estudo poderão servir de referência para trabalhos futuros semelhantes.

2.3.2 Métodos e Abordagens Propostas

Neste trabalho pretende-se não só incluir o estado passado das variáveis, tal como em [20, 21, 22], como também variáveis relativas a outras barragens, assim como [18]. Concretamente, pretende-se prever o caudal afluente de uma dada barragem, utilizando como base um conjunto de dados horários dessa barragem e da que se encontra imediatamente a montante. Para este fim, optou-se por estudar as séries temporais, como será descrito na subsecção 2.3.2.1. No entanto, este trabalho tem a particularidade acrescida de o foco ser a previsão em cheias, que por sua vez são acontecimentos raros e extremos. Para lidar com este facto, decidiu-se acrescentar ao estudo das séries temporais uma abordagem de regressão de quantis e de previsão de valores raros extremos, explicadas nas subsecções 2.3.2.2 e 2.3.2.3, respetivamente.

2.3.2.1 Séries Temporais

Uma série temporal é uma sequência de observações, no qual cada instância é obtida num instante específico no tempo. Geralmente, as séries temporais são discretas e periódicas, ou seja, são constituídas por um conjunto de amostras de uma dada grandeza, retiradas em instantes do tempo específicos e com o intervalo entre observações fixo. Este tipo de dados é extremamente comum, sendo a sua aplicação virtualmente infindável. Por exemplo, os dados mensais das vendas de uma empresa, os dados horários sobre as condições meteorológicas de uma dada zona ou um conjunto de variáveis monitorizadas regularmente de um processo industrial. De uma maneira geral, o objetivo de analisar este tipo de dados consiste em compreender o modelo e os mecanismos que influenciam uma dada série temporal e, se possível, prever valores futuros desta com base em dados históricos [24]. O modelo clássico pelo qual se pode decompor uma série temporal é constituído por três componentes [25]:

- Tendência: Comportamento de lenta variação, que indica a direção geral que uma série temporal toma ao longo do tempo.
- Sazonalidade: Variações que se repetem ao longo do tempo, de uma forma periódica ou calendarizada.
- Ruído estacionário: Acontecimentos esporádicos.

Na figura 2.4, podemos observar a decomposição gráfica de uma série temporal nestas três componentes.

Uma abordagem à previsão de séries temporais consiste em remover a influência da tendência e do comportamento sazonal, ficando apenas com a componente esporádica, e criar um modelo capaz de prever essa terceira componente. No fim, as previsões são obtidas aplicando o modelo dos acontecimentos esporádicos e adicionando a tendência e a componente sazonal. Este método é indicado, por exemplo, para situações em que se cria o modelo com dados relativos a um determinado número de anos e se prevê os anos seguintes. Existem também técnicas que permitem obter um modelo único da série temporal, sem ser preciso separar as suas componentes, de entre as

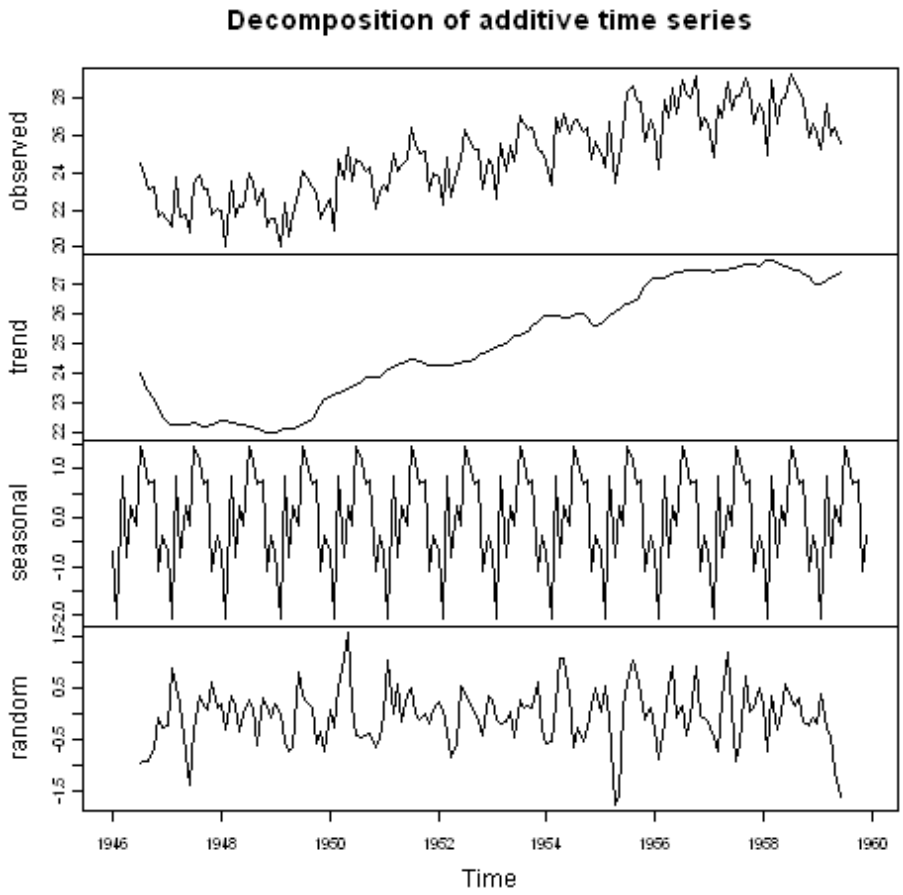


Figura 2.4: Decomposição de uma Série Temporal [26]

quais se destaca a regressão com valores atrasados, denominada modelo AR (*Autoregressive*) [27]. Este algoritmo, tenta encontrar um modelo linear que explique a dependência entre a variável a prever x_t e os seus valores passados $x_{(t-1)}, x_{(t-2)}, ..., x_{(t-n)}$, sendo que cada valor atrasado é tratado como uma variável independente. A equação 2.3 descreve matematicamente este algoritmo.

$$x_t = w_0 + w_1 x_{t-1} + ... + w_p x_{t-p} + \varepsilon_t \tag{2.3}$$

A primeira constante, w_0, representa a intersecção com o eixo, as restantes constantes em w são os coeficientes atribuídos a cada valor atrasado da variável a prever e ε_t é o erro que se pretende minimizar. Este modelo permite, face uma sequência de valores, prever o valor seguinte na série temporal. Ao número de valores passados utilizados para fazer a previsão, representado pela letra p, chama-se *lag* ou ordem do modelo e à quantidade de valores futuros a prever chama-se horizonte de previsão [27]. Observando a equação 2.3, conclui-se que, quando o horizonte de previsão é superior a um, faltam termos observados em x_{t-1}, x_{t-2} até x_{t-h+1}. Geralmente utiliza-se os valores previstos em cada iteração para substituir estes termos, passando as novas previsões a depender das anteriores. Outra forma de abordar o problema seria determinar uma equação para

cada valor do horizonte de previsão, que dependesse apenas dos valores passados [28]. Assim, em vez da equação 2.3 passaríamos a ter o conjunto de equações 2.4.

$$\begin{cases} x_t = w_{01} + w_{11}x_{t-1} + ... + w_{p1}x_{t-p} + \varepsilon_t \\ x_{t+1} = w_{02} + w_{12}x_{t-2} + ... + w_{p2}x_{t-p+1} + \varepsilon_{t+1} \\ x_{t+h-1} = w_{0h} + w_{1h}x_{t-h} + ... + w_{ph}x_{t-p+h-1} + \varepsilon_{t+h-1} \end{cases} \tag{2.4}$$

É de reparar que o conceito do tempo está implícito nas expressões anteriores através do uso dos valores atrasados em p, independentemente dos dados serem horários, mensais ou anuais, não sendo por isso necessário acrescentar variáveis correspondentes ao tempo

Todas estas abordagens aqui apresentadas consideram que a série temporal é univariada, ou seja, o valor futuro da série depende apenas dos valores passados de si própria. Nestes casos, existe apenas uma variável contínua que está a ser registada periodicamente e a qual queremos prever. No entanto, em muitas aplicações reais, uma série temporal não depende apenas de si própria, sendo normalmente influenciada por outras variáveis que, por sua vez, podem também ser consideradas séries temporais. Assim, obtém-se uma série temporal multivariada, onde cada variável corresponde a uma série que depende dos seus valores passados e dos valores passados das outras. É exemplo de uma série temporal multivariada o conjunto da pressão, temperatura e volume de um líquido ao longo do tempo [29].

As séries temporais são um método interessante para estudar o comportamento das variáveis hidrológicas de uma barragem pois, não só os dados são periódicos como o tempo é um fator muito importante, pretendendo-se fazer previsões para várias horas.

2.3.2.2 Regressão de Quantis

Os quantis de uma dada variável representam as fronteiras entre partições da sua distribuição, possuindo cada uma aproximadamente o mesmo número de elementos. Uma instância pertence ao quantil q se for superior à instância de referência correspondente à porção q e for inferior à de $1-q$ [30]. A probabilidade de uma variável aleatória ser menor que dado um quantil q é, aproximadamente, q/n e a probabilidade de ser maior $1-q/n$, em que n é o número de quantis. De entre os quantis mais comuns pode-se salientar os quartis, para $n = 4$, e os percentis, para $n = 100$ [31]. A figura 2.5 apresenta a distribuição de um conjunto de dados em quartis e a posição de cada um destes. Em seguida apresenta-se também um exemplo prático [32].

Considere-se as alturas em centímetros de cinco pessoas: $150, 200, 180, 160, 170$. O primeiro passo na determinação dos quantis é ordenar a sequência, obtendo-se: $150, 160, 170, 180, 200$. O valor do q^{o} quartil é dado pelo valor da sequência ordenada na posição $\frac{q}{4}(N+1)$, em que N é o número de valores na sequência, sendo neste caso igual a cinco. Por exemplo, o 1^{o} quartil seria o valor na posição 1.5, ou seja, $\frac{150+160}{2} = 155$. Segundo este conjunto de dados, a probabilidade de uma pessoa medir menos de $155cm$ é de 25%.

Os quantis permitem então analisar a distribuição dos dados relativamente à sua frequência e probabilidade de ocorrência. No entanto, os métodos de regressão linear procuram fazer uma

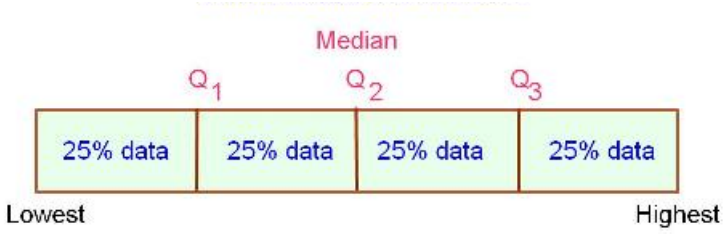


Figura 2.5: Distribuição de um conjunto de dados em quartis [33]

aproximação média [15], não sendo por isso capazes de modelar variações entre quantis. Assim, utilizando a regressão de quantis, é possível criar de modelos que, em vez de aproximarem todas as observações por uma mesma função, têm em consideração as diferenças entre quantis, podendo criar diferentes aproximações para cada um. Por exemplo, uma forma simples de implementar a regressão de quantis consiste em criar um modelo linear para cada quantil, originando um conjunto de coeficientes para cada um destes [34]. Por sua vez, cada coeficiente traduz a influência de uma variável independente na variável dependente, para o quantil em questão. Dado que a regressão linear tenta minimizar a soma quadrática dos resíduos, pode-se esperar que a solução se aproxime do comportamento médio, ao passo que, se se minimizar a soma absoluta dos resíduos, esta irá se aproximar mais da mediana. Este raciocínio é generalizável a outros quantis, bastando apenas atribuir diferentes pesos aos resíduos. Matematicamente, considerando que para qualquer que seja o quantil, $0 < q < 1$, com uma *loss function* como a ilustrada na figura 2.6, o objetivo torna-se resolver a equação 2.5 [30], em que q se encontra representado pela letra grega τ.

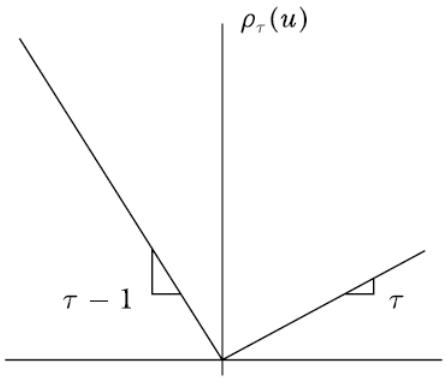


Figura 2.6: *Loss function* utilizada na regressão de quantis [30]

$$\min_{\beta \in \Re^p} \sum \rho_\tau (y_i - \varepsilon(x_i, \beta)) \tag{2.5}$$

Nesta equação, x_i e y_i representam respetivamente o conjunto de dados de entrada e saída do modelo, e ε é uma função linear.

A regressão de quantis é especialmente vantajosa quando existe uma gama da variável dependente mais relevante que os restantes valores, pois permite a criação de um modelo direccionado para um dado quantil que seja mais representativo dessa gama. Ao indicar o quantil em que se pretende focar, o modelo é otimizado para prever corretamente esses casos, à custa de um pior desempenho nos restantes [35]. Para além disso, a regressão de quantis permite estudar a variação da importância que cada variável independente tem na variável dependente, para diferentes quantis. Tomando como exemplo a tabela presente na figura 2.7, que faz parte de um estudo sobre o impacto de vários factores no peso de um bebé a quando o nascimento, é possível concluir que, à medida que a variável peso vai diminuindo, a variável cuidados pré-natais vai aumentando de importância [34].

Characteristic	**Linear Regression**	**Quantile Regression**				
		5th	**10th**	**50th**	**90th**	**95th**
Intercept	3224	2353	2608	3252	3856	4031
Married	161.1	227	171	149	141	165
Boy	115.9	28	84	121	142	142
Prenatal care	-227.0	-536	-418	-164	-111	-57
Smoke	-200.9	-255	-226	-190	-177	-199

Figura 2.7: Tabela comparativa entre a regressão linear e a regressão de quantis na previsão do peso de nascimento [34]

Esta abordagem é então relevante para a previsão de variáveis hidrológicas em épocas de cheias pois, sabendo que estas tomarão valores extremos, é possível orientar o modelo para prever corretamente estes casos, visto serem os mais relevantes, à custa de um pior desempenho nos outros.

2.3.2.3 Previsão de valores raros extremos

A previsão de valores raros extremos procura criar modelos adaptados para prever corretamente os casos mais raros, à custa de uma pior previsão nos casos comuns [36]. Esta tarefa é particularmente difícil pois, para além dos casos raros terem uma frequência de ocorrência menor, tornando mais difícil a identificação de padrões, as medidas *standard* de performance não são aplicáveis, dado o comportamento geral não ser importante mas sim o desempenho nos casos raros. Este estudo é particularmente útil para aplicações em que uma situação normal é pouco relevante, como por exemplo, em sistemas destinados a detetar falhas ou situações de perigo. Neste caso a variável dependente é discreta e toma valores binários correspondentes a uma situação normal ou de anomalia e será muito mais frequente ocorrer uma situação normal do que o contrário, apesar de ser mais importante detetar as falhas. Assim, um modelo que minimize o erro médio poderá, no limite, prever sempre que a situação é normal, obtendo uma boa performance global mas sendo completamente inútil para a aplicação em questão.

Para modelos de classificação, como no exemplo dado, este problema está vastamente estudado, tendo sido propostas várias soluções. Nomeadamente, uma maneira muito simples de

resolver este problema, quando se sabe quais as classes mais importantes, é atribuir diferentes custos aos erros [37], consoante a classe em questão. Este método, chamado *Cost-Sensitive Learning*, permite criar modelos que, em vez de minimizarem o erro médio, minimizam o custo total, associado aos custos parciais atribuídos para cada previsão. De uma forma geral, relativamente à classificação, pode-se enumerar três abordagens mais comuns para resolver a previsão de valores raros extremos [36]:

- Alterar o método de avaliação, tal como no exemplo apresentado anteriormente.
- Modificar o algoritmo a implementar, por forma a adaptar o processo de otimização aos objetivos do problema.
- Aplicar técnicas de amostragem que alterem a distribuição dos dados de treino, permitindo que se possa aplicar a algoritmos e métodos de avaliação *standard*.

Apesar destas técnicas serem adaptáveis para o caso de variáveis contínuas, não existe tanta informação e estudos disponíveis, aumentando a dificuldade de implementação. Dos estudos que existem disponíveis, pode-se salientar um que altera o método de avaliação [38] e outro que aplica técnicas de amostragem [36]. O primeiro, define um método de avaliação que valoriza a previsão dos casos raros, consistindo basicamente numa média pesada dos erros, enquanto que o segundo implementa um método de sub-amostragem, diminuindo o número de amostras das observações consideradas normais, e uma adaptação do método SMOTE (*Synthetic Minority Over-sampling Technique*) que, por sua vez, utiliza simultaneamente sub-amostragem dos casos frequentes e sobre-amostragem dos casos raros.

Pode-se observar que este método tem um propósito semelhante à regressão de quantis mas, ao passo que esta permite adaptar o modelo para uma dada gama da variável dependente, preocupando-se mais com o valor que ela toma do que com a sua frequência, a previsão de valores raros procura criar modelos para os casos menos frequentes, podendo esses tomar valores extremos ou não. Geralmente, os valores raros são também valores extremos, daí a sua dificuldade de previsão, como é o caso do caudal afluente em cheias pois espera-se que tome valores mais altos e que a sua frequência anual seja baixa. É mais comum uma situação normal, em que o controlo dos descarregamentos está bem definido, do que o contrário. Portanto, dado que se pretende fazer previsões em situações menos frequentes, a previsão de valores raros extremos é também um método interessante para a resolução deste problema.

2.4 Resumo e Conclusões

Nos trabalhos referidos na subsecção 2.3.1, podemos notar que existe efetivamente uma preocupação em estudar e otimizar a Gestão de Descarregamentos. Apesar da maioria dos estudos ter como objetivo final a criação de modelos para previsão do caudal a ser descarregado num dado momento, todos contribuem grandemente para a sua compreensão.

Esta tese propõe uma nova abordagem, que não pretende prever diretamente os descarregamentos mas sim algumas das variáveis que os condicionam, de forma a permitir ao operador tomar decisões com base não só nas informações presentes, mas também com as estimativas futuras. Para além disso, propõe-se a utilização de dados horários em conjunto com dados esporádicos, o que corresponde a uma granularidade muito mais fina à utilizada noutros trabalhos.

Segundo as categorias referidas na secção 2.3, podemos descrever a abordagem proposta da seguinte forma:

- Tarefa de *Data Mining*: Previsão numérica;
- Algoritmos a utilizar: VAR (*Vector Autoregressive*), SZBVAR (*Reduced form Sims-Zha Bayesian VAR*), VMA (*Vector Moving Average*), DYNLM (*Dynamic Linear Model*), *random forests* (RF), redes neuronais (NN) e *conditional inference random forests* (CRF);
- Variável dependente: Caudal afluente;
- Variáveis independentes: Caudal afluente, caudal descarregado, cota montante, etc; São incorporados valores passados e variáveis da barragem imediatamente a montante;
- Granularidade dos dados: Horários e esporádicos;
- Método utilizado para avaliar a performance dos modelos: RMSE (*Root Mean Squared Error*) e MAPE (*Mean Absolute Percentage Error*).

A escolha dos algoritmos e o seu funcionamento será explicado na secção 4.1.

Capítulo 3

Compreensão dos dados

Ao longo deste capítulo apresenta-se uma análise exploratória realizada sobre os dados, de modo a averiguar algumas características relevantes das variáveis a trabalhar. Na secção 3.1 encontra-se o dicionário de dados, com o nome de cada variável e o seu significado para o problema, e na secção 3.2 as estatísticas básicas, incluindo a análise dos histogramas e gráficos de dispersão.

3.1 Dicionário de dados

O dicionário de dados permite averiguar quais as variáveis disponíveis ao estudo, ao mesmo tempo que fornece uma descrição para cada uma de forma a obter-se uma noção geral do conjunto de dados a tratar. A tabela 3.1 apresenta o dicionário de dados deste trabalho e está dividida em três colunas, correspondendo ao nome que cada variável toma na base de dados, o seu nome real e o seu significado. As variáveis disponíveis são as mesmas em ambos os conjuntos de dados de Venda Nova e Alto Rabagão, variando apenas a gama de valores que estas tomam. O código e sigla do centro são únicas a cada conjunto de dados mas, para o estudo em questão, são variáveis dispensáveis pois o seu valor é constante. Optou-se por, ao fazer a integração destas variáveis, renomeá-las de forma a manter o significado original mas acrescentando uma letra informativa da barragem. Para os dados de Venda Nova acrescentou se um *J*, indicando que se tratava da barragem mais a jusante no estudo e para Alto Rabagão um *M* de montante. Desta forma, permite-se uma abstração do conjunto específico de barragens, tornando todo o trabalho realizado aplicável a qualquer conjunto de duas barragem em cascata.

Tabela 3.1: Significado das variáveis horárias disponíveis

Campo	Nome da variável	Descrição
CENT_APROV_CENTCOD	Código do centro	Código identificador do centro de produção.
CENT_APROV_SIGLA	Sigla do centro	Sigla atribuída ao centro de produção.
CENT_DATA	Data	Data em que foram recolhidas as medidas. É apresentada no formato ano.mês.dia.
CENT_HORA	Hora	Hora em que foram recolhidas as medidas, de 0 a 23.
CENT_COTAMONT	Cota montante	Nível da água, em m, na parte montante da barragem (albufeira).
CENT_COTAJUS	Cota jusante	Nível da água, em m, na parte jusante da barragem.
CENT_VOLUME	Volume	Volume de água armazenado na albufeira, em hm^3.
CENT_CAUDTURB	Caudal turbinado	Quantidade de água que passa de montante para jusante da barragem, em m^3/s, no processo de produção de energia.
CENT_CAUDAFLUT	Caudal afluente	Quantidade de água que chega a montante da barragem, em m^3/s.
CENT_CAUDDESC	Caudal descarregado	Quantidade de água que passa de montante para jusante da barragem, em m^3/s, e que não é aproveitada para produzir energia, servindo apenas para libertar a água em excesso.
CENT_CAUDTRANSF	Caudal transferido	Quantidade de água que é transferida de uma barragem para outra, em m^3/s, por via de condutas próprias.
CENT_CAUDECOLOG	Caudal ecológico	Quantidade de água que passa de montante para jusante da barragem, em m^3/s, por via de um descarregador próprio para o efeito, permitindo garantir o caudal mínimo do rio.
CENT_CAUDBOMB	Caudal bombado	Quantidade de água que passa de jusante para montante da barragem, em m^3/s, através da bombagem da água. A água bombada pode ser posteriormente utilizada para produzir energia.

3.2 Estatísticas básicas

Os dados relativos às variáveis horárias registadas para Venda Nova e Alto Rabagão foram recolhidos da base de dados pela EDP e transferidos para folhas excel. Usou-se o MATLAB (MATLAB 6.1, The MathWorks Inc., Natick, MA, 2000) para importar esses dados, convertendo as células sem conteúdo em zeros. Em seguida, também no MATLAB, elaborou-se uma série de estatísticas básicas e gráficos que permitem compreender melhor a natureza e distribuição das variáveis em questão, utilizando dados relativos aos anos de 2001 a 2004.

Em primeiro lugar, retirou-se o máximo, o mínimo e a média de cada variável, de forma a obter uma noção da gama de valores que cada uma destas toma. Calculou-se também o 3º quartil, utilizando o comando *"quantile"* do MATLAB, de modo a adquirir maior sensibilidade relativamente aos valores elevados de cada variável, visto o máximo poder ser uma ocorrência isolada mas os valores superiores ao 3º quartil representarem os 25% valores mais altos da sua distribuição. Esta informação poderá ser útil dado que as cheias corresponderão aos valores mais altos das variáveis hidrológicas. A tabela 3.2 contém estas informações relativas ao conjunto de dados horários de Venda Nova e a tabela 3.3 apresenta as mesmas informações sobre os dados de Alto Rabagão.

Tabela 3.2: Estatísticas básicas sobre os dados horários de Venda Nova

Nome da variável	Mínimo	Máximo	Média	3º Quartil
Cota montante	662.55	700.46	690.94	698.60
Cota jusante	28.20	286.20	286.19	-
Volume	11.83	96.22	69.68	89.27
Caudal turbinado	0	35.04	13.04	28
Caudal afluente	-12.63	695.99	18.6179	27.18
Caudal descarregado	0	701.04	3.7659	0
Caudal transferido	0	0	-	-
Caudal ecológico	0	0	-	-
Caudal bombado	0	0	-	-

Tabela 3.3: Estatísticas básicas sobre os dados horários de Alto Rabagão

Nome da variável	Mínimo	Máximo	Média	3º Quartil
Cota montante	862.89	880.16	872.02	877.52
Cota jusante	0	700.46	697.01	698.61
Volume	261.93	572.52	413.85	512.73
Caudal turbinado	0	130.58	10.10	20.43
Caudal afluente	0	387.11	8.78	2.59
Caudal descarregado	0	202.93	0.28	0
Caudal transferido	0	0	-	-
Caudal ecológico	0	0	-	-
Caudal bombado	0	39.63	1.84	0

Analisando a tabela 3.2 e a informação recolhida na documentação relativa às normas de Venda Nova [8], fornecidas pela EDP, foi possível retirar algumas conclusões sobre a validade dos valores apresentados.

- Os valores de cota montante estão dentro do esperado, visto as normas estabelecerem que esta esteja sempre entre os $610,7m$ e $700,50m$.
- O valor mínimo de cota jusante é muito suspeito, não só por ser dez vezes mais pequeno que a média mas também por exceder largamente o limite mínimo imposto de $286,20m$.
- Convertendo os limites relativos à cota montante, através das tabelas de conversão de cota montante para volume de Venda Nova, que podem ser consultadas no Anexo A.3, obteve-se que este último se deve encontrar entre os $0m^3$ e $96.37m^3$. Mais uma vez, esta informação vai de encontro aos valores recolhidos.
- Observa-se que o caudal turbinado excede um pouco o valor máximo definido de $30m^3/s$.
- O valor mínimo do caudal afluente é claramente um erro pois, dado o conhecimento fornecido pelos peritos no domínio, este não pode tomar valores negativos.
- Para o caudal descarregado existe um limite máximo de $1100m^3/s$, que está acima do valor máximo obtido, significando que nunca foi necessário descarregar ao máximo.
- Os zeros no transferido, ecológico e bombado podem significar que a barragem de Venda Nova não possui infraestruturas para realizar estes descarregamentos ou, se as possui, não as utilizou no período observado.

A mesma análise pode ser feita a partir da tabela 3.3 e da informação recolhida na documentação fornecida pela EDP [5], retirando-se as seguintes conclusões relativas aos dados de Alto Rabagão:

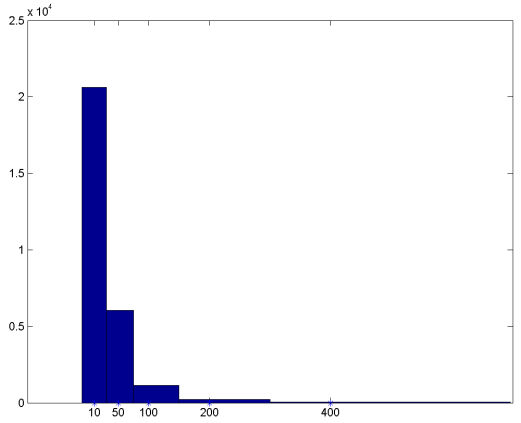
- Os valores da cota montante estão dentro dos limites estabelecidos pela norma, sendo estes o valor mínimo de $798.60m$ e máximo de $880.10m$.
- Mais uma vez, a cota jusante apresenta um valor mínimo suspeito, dado ser muito mais pequeno que a média e exceder o valor mínimo imposto de $695.00m$.
- O volume está de acordo com os valores obtidos a partir da tabela dos volumes em função da cota. Fazendo a conversão de cota para volume obtém-se um mínimo de $261.92m^3$ e máximo de $571.08m^3$. Mais uma vez, as tabelas de conversão cota-volume de Alto Rabagão podem ser consultadas no Anexo A.3
- O valor máximo de caudal turbinado excede o valor definido de $47m^3/s$. Uma quantidade significativa de dados excede este valor fica a baixo de $64m^3/s$ e apenas dois pontos são superiores a $100m^3/s$.

- O valor máximo de caudal descarregado está dentro dos limites estabelecidos, sendo o valor máximo imposto de $860.00 m^3/s$.
- O caudal bombado máximo excede o limite definido de $37 m^3/s$.
- Tal como em Venda Nova, os zeros no transferido e ecológico podem significar que a barragem de Alto Rabagão não possui infraestruturas para realizar estes descarregamentos ou, se as possui, não as utilizou no período observado.

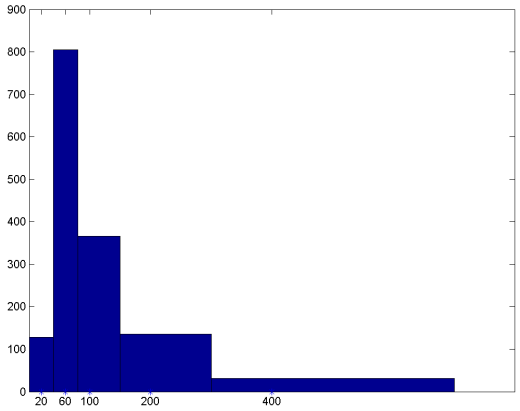
3.2.1 Histogramas

Os histogramas permitem obter uma distribuição aproximada de uma dada variável contínua. Para tal, partindo de um conjunto de dados observados dessa variável, agrupam-se os exemplos em partições e determina-se o número de ocorrências em cada partição [39]. O gráfico obtido através desta análise fornece informações importantes sobre a variável em si e sobre o conjunto de dados recolhidos, podendo-se até mesmo detetar *outliers*. Estes são valores que não estão de acordo com a distribuição normal dos dados [15]. Numa primeira fase observou-se que os caudais afluentes, descarregado e turbinado apresentam maioritariamente o valor zero, tendo por isso uma coluna dominante no histograma que dificulta a visualização dos restantes valores. Dado que em cheias esperar-se que estas variáveis não sejam nulas, optou-se por representar os seus histogramas sem os zeros. A figura 3.1 contém todos os histogramas obtidos para as variáveis de Venda Nova e a figura 3.2 os de Alto Rabagão.

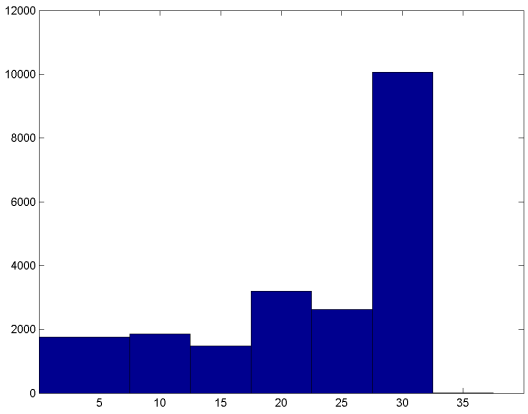
Como se pode observar pela primeira coluna do histograma da figura 3.1a, mesmo sem os zeros o caudal afluente apresenta maioritariamente valores baixos. Isto deve-se ao facto de, quando a barragem a montante não descarrega significativamente e não há chuva, o caudal afluente é muito baixo. Sendo assim, pode haver situações em que não se está a descarregar, nem a turbinar e mesmo assim não há aumento significativo de volume, devido ao facto de o caudal afluente ser praticamente nulo. Geralmente, nas situações de baixo afluente o caudal descarregado é zero ou muito baixo. Analisando o histograma 3.1b, observa-se que quando existem descarregamentos, o seu valor típico ronda os $60 m^3/s$. Na figura 3.1c observa-se que o caudal turbinado apresenta uma concentração mais ou menos constante nos valores baixos, sendo o pico da sua distribuição em $30 m^3/s$. Dado que os valores máximos do turbinado são bastante menores que os do descarregado, pode-se supor que esta variável vá ser de menor importância numa situação de cheia, em que os caudais são muito superiores a este, podendo o turbinado ser desprezável em relação ao descarregado. Da figura 3.1d conclui-se que a gama de valores da cota jusante é muito restrita, podendo-se considerar que esta variável é constante. Sendo assim, o valor mínimo apresentado de 28.20 é um *outlier* e esta variável pode ser excluída do conjunto de variáveis independentes, dado ser constante e igual a $286,20 m$. A cota montante, cujo histograma se pode observar na figura 3.1e, concentra-se maioritariamente em valores altos, tal como o volume que apresenta um histograma semelhante a este, como seria de esperar dado estes dois valores estarem directamente relacionados.



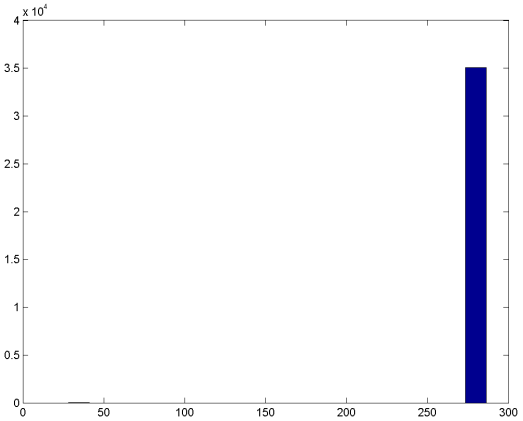
(a) caudal afluente, sem os zeros



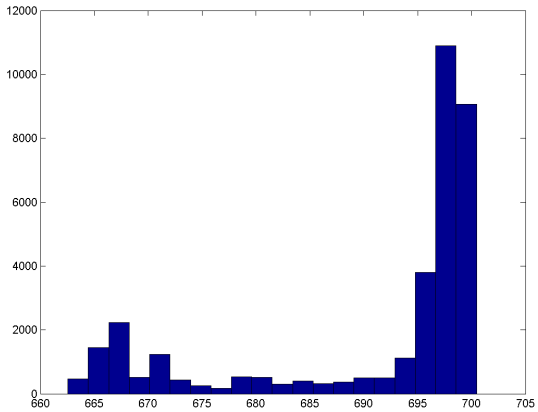
(b) caudal descarregado, sem os zeros



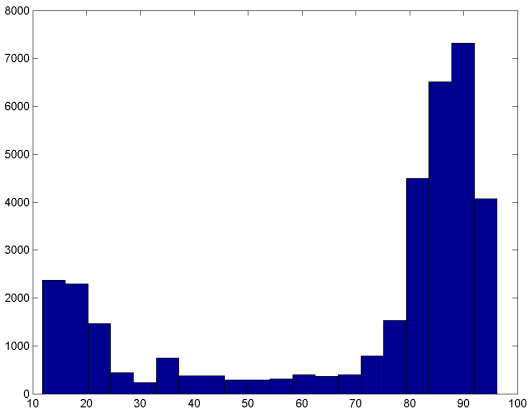
(c) caudal turbinado, sem os zeros



(d) cota jusante

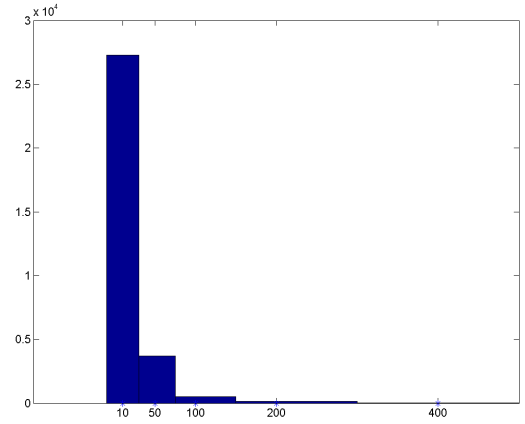


(e) cota montante

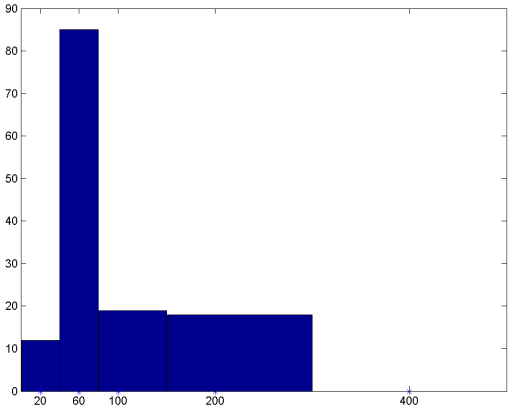


(f) volume

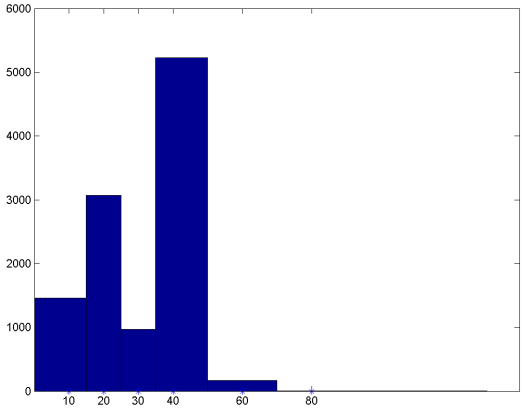
Figura 3.1: Histogramas obtidos para as variáveis horárias de Venda Nova



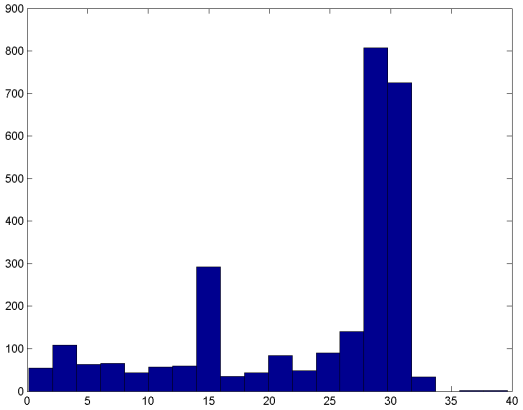
(a) caudal afluente, sem os zeros



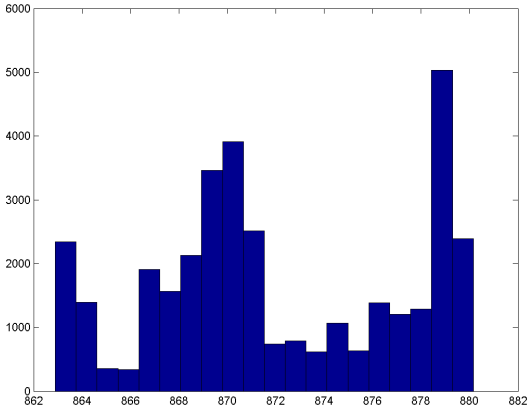
(b) caudal descarregado, sem os zeros



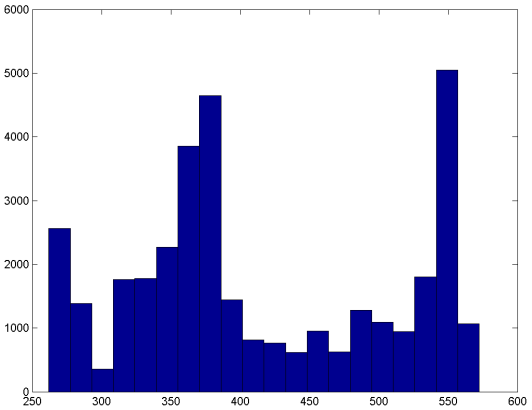
(c) caudal turbinado, sem os zeros



(d) caudal bombado, sem os zeros



(e) cota montante



(f) volume

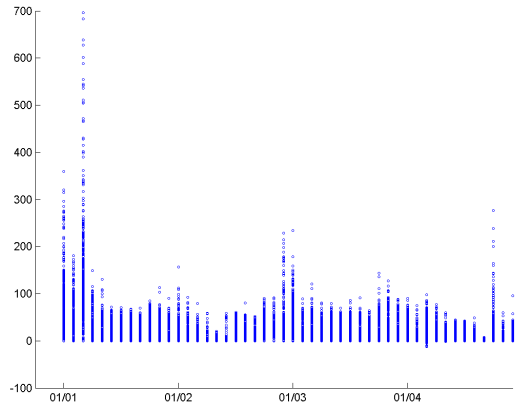
Figura 3.2: Histogramas obtidos para as variáveis horárias de Alto Rabagão

As conclusões a retirar dos histogramas relativos às variáveis de Alto Ragabão, apresentados na figura 3.2, são semelhantes ao que foi apresentado para Venda Nova. Mais uma vez, o caudal afluente apresenta maioritariamente valores baixos e o descarregamento mais frequente ronda os $60m^3/s$. No entanto, apesar da forma dos histogramas 3.1b e 3.2b ser semelhante, as frequências apresentadas não o são. Ao passo que, para Venda Nova o pico apresenta uma contagem de cerca de 800 pontos, para Alto Rabagão esta nem chega aos 90 pontos. Isto significa que Venda Nova descarrega muito mais que Alto Rabagão, sendo que nesta última obtém-se uma maior frequência de zeros. Para o caudal turbinado obtém-se uma conclusão análoga. Ao contrário de Venda Nova, Alto Rabagão possui bombagem mas, tal como as variáveis anteriores, esta é maioritariamente zero. Do histograma 3.2d, do caudal bombado sem os zeros, conclui-se que este apresenta uma concentração significativa em torno do valor 30. Novamente, a forma dos histogramas da cota montante, figura 3.2e, e do volume, figura 3.2f, são muito semelhantes entre si. Comparando com cota montante e volume de 3.1, as de 3.2 apresentam maior frequência de valores intermédios, não estando tão concentrados nos valores mais altos. Por último, optou-se por não apresentar o histograma da cota jusante que, apesar de não ser constante, apresenta uma gama de valores muito restrita.

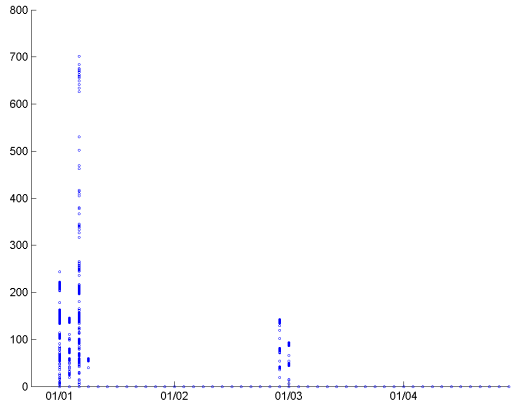
3.2.2 Gráficos de dispersão

Os gráficos de dispersão permitem observar a relação entre duas variáveis e possibilitam a identificação de padrões ou correlações entre estas [40]. Dado que se espera encontrar alguma sazonalidade nos dados, elaborou-se vários gráficos de dispersão dos dados em função do mês e do ano, ou seja, cada ponto no eixo das abcissas corresponde a um dado mês de um dado ano. Os gráficos de dispersão das variáveis de Venda Nova encontram-se representados na figura 3.3.

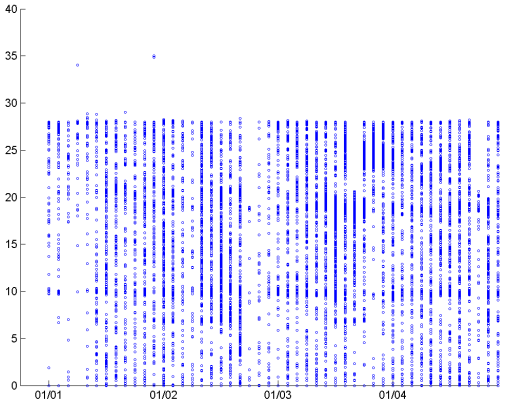
Observando o gráfico 3.3a, conclui-se que, tal como era esperado, existem algumas flutuações sazonais no caudal afluente de Venda Nova, sendo que por volta de Janeiro de cada ano há um aumento significativo deste. Essa variação é mais acentuada no início de 2001 devido à cheia excecional que ocorreu nesse ano [41]. Relativamente ao caudal descarregado, apresentado na figura 3.3b, observa-se novamente que este é maioritariamente baixo, tendo um pico em 2001 devido à cheia excecional. Este gráfico vem reforçar a ideia de que Venda Nova descarrega apenas ou maioritariamente em situações de cheia. Da figura 3.3c pode-se retirar que o caudal turbinado é praticamente independente da estação, podendo tomar qualquer valor entre o máximo e o mínimo. Dado que o turbinado é utilizado para produzir energia, faz sentido que este seja mais ou menos independente do mês ou ano que se está a considerar. Uma observação interessante é que durante a cheia de 2001 o caudal turbinado toma valores maioritariamente altos, o que leva a crer que devido ao excesso de água houve também um excesso de produção. A cota montante e o volume, que mais uma vez apresentam gráficos muito idênticos, têm um comportamento sazonal interessante. Como seria espectável, estes diminuem durante os meses de calor, podendo-se observar descidas acentuadas entre Maio e Outubro de 2001 e entre Junho e Outubro de 2002. Em 2003 esse efeito é muito mais suave e em 2004 faz-se apenas sentir entre Agosto e Outubro.



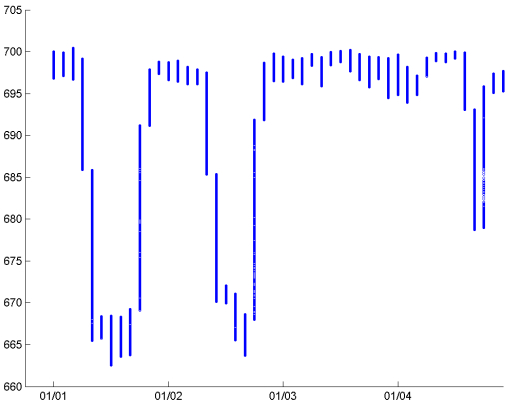
(a) caudal afluente



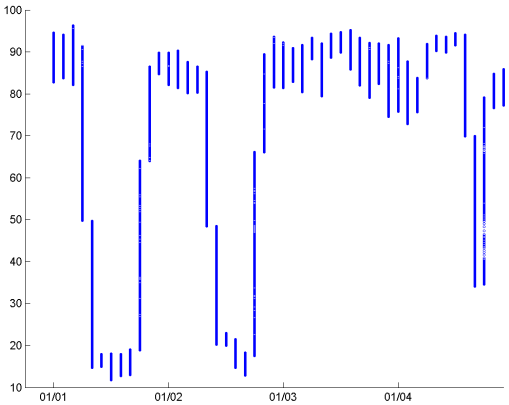
(b) caudal descarregado



(c) caudal turbinado

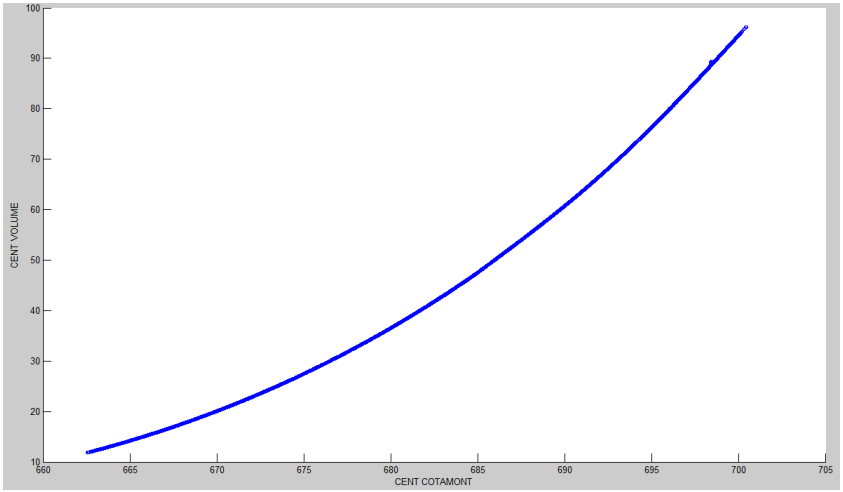


(d) cota montante

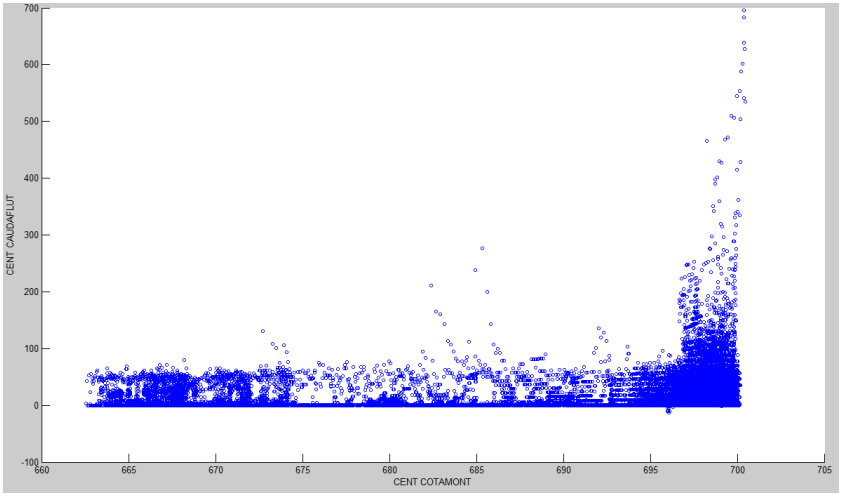


(e) volume

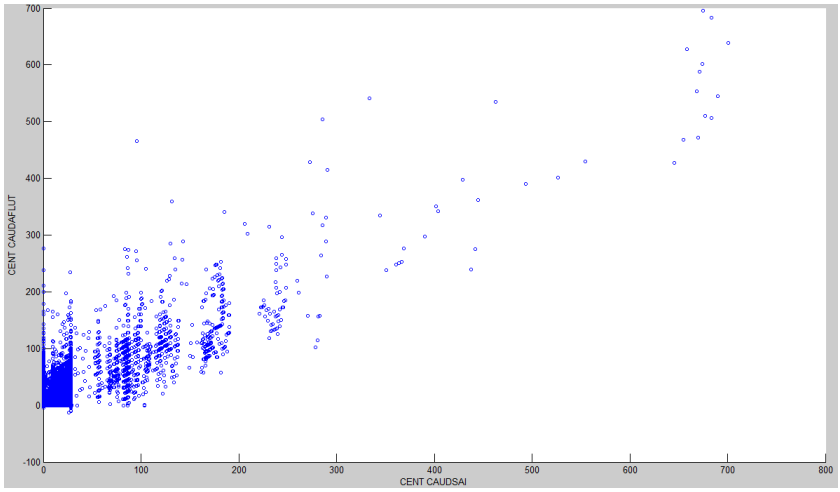
Figura 3.3: Gráficos de dispersão obtidos por ano e mês para as variáveis horárias de Venda Nova



(a) cota montante e volume

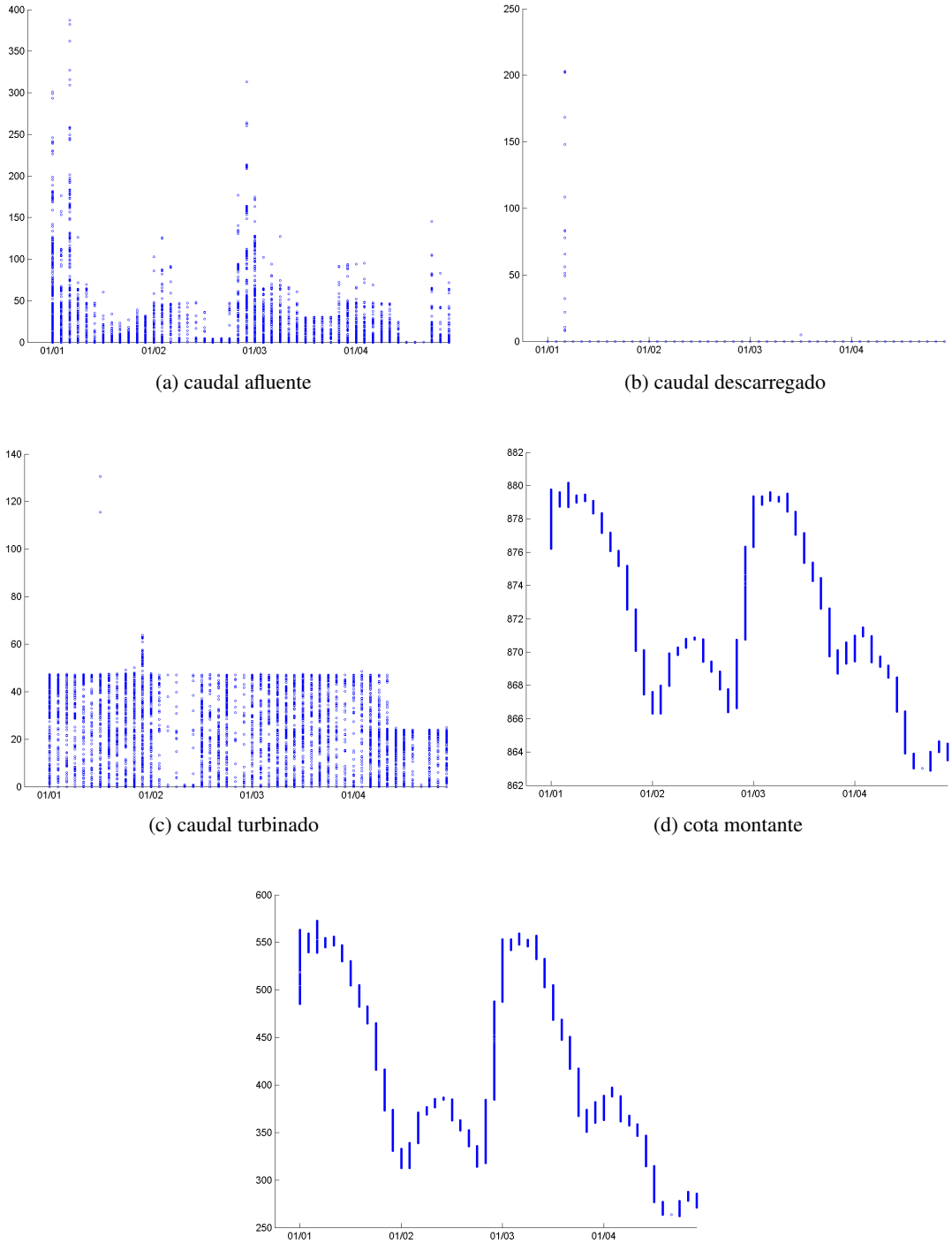


(b) caudal afluente e cota montante



(c) caudal afluente e caudal libertado

Figura 3.4: Gráficos de dispersão entre pares de variáveis de Venda Nova



(a) caudal afluente

(b) caudal descarregado

(c) caudal turbinado

(d) cota montante

(e) volume

Figura 3.5: Gráficos de dispersão obtidos por ano e mês para as variáveis horárias de Alto Rabagão

Para além dos gráficos de dispersão em função do tempo, elaboraram-se alguns gráficos para pares de variáveis que poderão ter algum tipo de correlação, estando estes representados na figura 3.4. Em primeiro lugar, o gráfico 3.4a apresenta a relação entre a cota montante e o volume da albufeira de Venda Nova que, tal como se esperava, possuem uma correlação positiva quase perfeita. Aliás, o coeficiente de correlação entre estas duas variáveis é de 0.9933, comprovando que uma é directamente proporcional à outra. Assim, uma destas variáveis pode ser removida do conjunto de dados a trabalhar, pois não é necessário ter duas variáveis linearmente dependentes. Analisando a figura 3.4b conclui-se que não existe nenhuma correlação aparente entre o caudal afluente e a cota montante, no entanto, pode-se retirar que valores altos de caudal afluente implicam sempre valores altos de cota montante (o contrário já não é válido). Por último, na figura 3.4c observa-se que existe uma correlação positiva entre o caudal afluente e o libertado em Venda Nova, sendo o coeficiente de correlação 0.7457. Isto seria de esperar dado que $Q_{afluente} = Q_{turbinado} + Q_{descarregado} \pm \frac{\Delta V}{\Delta t}$.

As conclusões a retirar relativamente a Alto Rabagão, a partir dos gráficos presentes na figura 3.5, são muito semelhantes às apresentadas para Venda Nova. A única diferença significativa manifesta-se no comportamento do caudal turbinado, cota montante e, consequentemente, volume em grande parte do ano 2004. Em 3.5c pode-se observar uma queda abrupta nos últimos meses, acompanhada de uma diminuição acentuada da cota montante e volume na albufeira de Alto Rabagão. Este fenómeno é possivelmente causado pela seca que ocorreu entre 2004 e 2005, devido à falta de chuva em Portugal Continental[42] que por consequência provoca uma diminuição no saldo de água disponível nas barragens. Dado que Alto Rabagão é a primeira barragem na cascata do Cávado-Lima, este fenómeno é mais acentuado nesta e, devido à sua retenção de água, é atenuado nas barragens a jusante desta.

3.3 Casos de cheia

Como referido várias vezes ao longo deste relatório, o objetivo é fazer previsões em cheias. Segundo a informação fornecida pela EDP, sempre que a barragem de Alto Rabagão estiver a descarregar estamos perante um cenário de cheia. Isto deve-se ao facto de a sua albufeira ter um papel de acumulação na cascata do Cávado-Lima, sendo portanto a água libertada por esta preferencialmente pela forma de caudal turbinado. Para evitar alguns erros cometidos pela existência de valores perto do zero, dado os sensores não serem perfeitos, considerou-se que existe uma cheia sempre que o descarregado de Alto Rabagão for superior a $25 m^3/s$, valor estabelecido pelos peritos do domínio na EDP.

3.3.1 Estatísticas em cheia

Decidiu-se também elaborar um pequeno estudo estatístico às cheias presentes em 2001, na expectativa de retirar informações relevantes para o desenvolvimento do trabalho. A tabela 3.4 e 3.5 contêm as mesmas medidas que as tabelas 3.2 e 3.3, mas para o caso de cheia em Venda Nova e Alto Rabagão, respetivamente. A partir destas tabelas, conclui-se que, como era de esperar, os valores mínimos e médios das variáveis subiram, sendo essa subida mais acentuada para o caudal afluente e descarregado. A única excepção à regra é o caudal bombado, que deixa de existir em caso de cheia para Alto Rabagão. Ou seja, em caso de cheia existe água em excesso e portanto não faz sentido bombar água para posterior acumulação.

Tabela 3.4: Estatísticas básicas sobre os dados horários de Venda Nova, em cheias

Nome da variável	Mínimo	Máximo	Média
Cota montante	696.67	700.46	697.87
Volume	82.18	96.22	86.57
Caudal turbinado	0	28	24.72
Caudal afluente	127.13	695.99	242.22
Caudal descarregado	95.66	701.04	239.47

Tabela 3.5: Estatísticas básicas sobre os dados horários de Alto Rabagão, em cheias

Nome da variável	Mínimo	Máximo	Média
Cota montante	879.81	880.10	879.93
Cota jusante	696.67	700.46	697.87
Caudal turbinado	44	47	46.94
Caudal afluente	13.07	382.47	112.42
Caudal descarregado	32.26	202.93	76.73
Caudal bombado	0	0	-

Ao observar os valores da cota montante de Venda Nova e a cota jusante de Alto Rabagão, chegou-se à conclusão que, em cheias, estas são idênticas. Aliás, somou-se a diferença entre estes dois conjuntos de dados horários e esta deu aproximadamente zero. Isto pode ser explicado pelo facto de a água em cheia se propagar com maior velocidade de uma barragem para a outra, podendo demorar menos de uma hora. Sugere-se então que a granularidade disponível pode não ser suficientemente fina para o estudo a realizar.

3.4 Limpeza dos dados

Na secção 3.2, referiu-se que algumas variáveis apresentam valores máximos e mínimos que excedem aqueles definidos pela documentação fornecida pela EDP. Considerando que valores fora destes limites se encontram fora do domínio da variável, um primeiro passo na preparação dos dados passa por corrigir esses valores. Optou-se então por realizar uma limpeza com base no conhecimento do domínio, obtido através das normas [8, 5], e validada pelos peritos. O código de

limpeza, realizado em R [43], iguala todos os valores que excedem a sua gama ao limite considerado, quer este seja um máximo ou um mínimo. Esta limpeza foi realizada apenas sobre os dados de modelação, tendo sido os dados de teste deixados em bruto, como seria numa situação real.

3.5 Resumo e conclusões

Ao longo deste capítulo foi possível explorar as características das variáveis hidrológicas disponíveis para o estudo, obtendo-se assim diversas conclusões úteis para o trabalho a realizar. Na secção 3.2, observou-se a gama de valores que cada uma das variáveis toma, concluindo-se que na sua maioria estas cumprem com o estabelecido pelas normas. Às exceções encontradas, optou-se por realizar uma limpeza com base no conhecimento do domínio, como explicado na secção 3.4. Ao elaborar e analisar os histogramas apresentados na subsecção 3.2.1, obteve-se algumas informações relativas à distribuição das variáveis importantes para o estudo das cheias, nomeadamente o facto de o caudal afluente, descarregado e turbinado tomarem maioritariamente valores iguais ou perto de zero. Ou seja, para além das cheias serem raras, obter comportamentos semelhantes aos observados em cheia, mesmo que não se enquadrem na definição apresentada na secção 3.3, é também pouco frequente. Através dos gráficos de dispersão presentes na subsecção 3.2.2, foi possível comprovar a existência de um comportamento sazonal, provocado pelas estações de cheia e seca, salientando as anomalias de 2001 e 2004 provocadas por fenómenos extremos de cheia e seca, respectivamente. Para além destas informações foi também possível concluir que algumas variáveis possuem características específicas que permitem excluí-las do estudo. Com base neste conhecimento decidiu-se retirar a cota jusante de Venda Nova, por ser constante, o volume de Venda Nova e Alto Rabagão por ser linearmente dependente da cota montante, o caudal transferido, ecológico e bombado de Venda Nova por ser constante e igual a zero e o caudal transferido e ecológico de Alto Rabagão pelos mesmos motivos anteriores.

No fim desta análise, obteve-se um conjunto final de dez variáveis sobre as quais o estudo irá incidir:

- Cota montante de Venda Nova
- Caudal afluente de Venda Nova
- Caudal descarregado de Venda Nova
- Caudal turbinado de Venda Nova
- Cota montante de Alto Rabagão
- Cota jusante de Alto Rabagão
- Caudal afluente de Alto Rabagão
- Caudal descarregado de Alto Rabagão

- Caudal turbinado de Alto Rabagão
- Caudal bombado de Alto Rabagão

As variáveis relativas ao tempo estão implícitas no conceito de série temporal, como foi referido na subsecção 2.3.2.1, não sendo por isso necessário incluir a data nem a hora no conjunto de variáveis a modelar.

Capítulo 4

Gestão de Descarregamentos: Abordagem de Previsão

Neste capítulo faz-se uma descrição da metodologia seguida e resultados obtidos relativamente à abordagem de previsão, dividindo-se este estudo em seis secções. Na secção 4.1 faz-se uma breve introdução aos algoritmos a utilizar e na secção 4.2 apresenta-se os resultados obtidos na primeira iteração. Na secção 4.3 apresenta-se os resultados da inclusão de novas variáveis nos modelos previamente referidos e na secção 4.4 descreve-se uma solução para tornar a granularidade dos dados mais fina, apresentando os respectivos resultados. Na secção 4.5 e 4.6 apresentam-se outras abordagens, tomando foco na regressão de quantis e na previsão de valores raros extremos, respectivamente. Por último, na secção 4.7, reúnem-se as conclusões obtidas ao longo deste capítulo.

4.1 Algoritmos e Metodologia

O conjunto de dados a utilizar consiste numa série de registos horários, em que cada variável corresponde à média horária de várias observações recolhidas ao longo daquela hora, sendo por isso uma série temporal multivariada de intervalo horário. Através do conhecimento no domínio do fenómeno, intui-se que haja uma componente sazonal anual forte, relacionada com as épocas de cheia e de secas, e possivelmente alguma sazonalidade diária relacionada com os períodos de descarga. No entanto, dado o objetivo ser fazer previsões de muito curto prazo, como por exemplo prever as próximas cinco horas, optou-se por seguir a segunda abordagem apresentada na subsecção 2.3.2.1 e tentar relacionar os próximos valores do caudal afluente com os valores imediatamente passados deste e das outras variáveis.

O modelo AR aí descrito pode ser extendido às séries temporais multivariadas, passando a denominar-se modelo VAR. A equação 4.1 descreve um modelo VAR de ordem 1 com duas séries, $x_{t,1}$ e $x_{t,2}$ [44].

$$x_{t,1} = w_{01} + w_{11}x_{t-1,1} + w_{12}x_{t-1,2} + \varepsilon_{t,1} \qquad x_{t,2} = w_{02} + w_{21}x_{t-1,1} + w_{22}x_{t-1,2} + \varepsilon_{t,2} \tag{4.1}$$

A interpretação das constantes é semelhante ao que foi referido para a equação 2.3, a única diferença significativa é o facto de cada série depender também dos valores passados das outras séries. Como referido anteriormente, na maioria dos casos a regressão é feita minimizando o erro ε pelo método dos mínimos quadrados. No entanto, existem algumas variantes do modelo VAR que utilizam outros métodos de estimação de parâmetros, como o SZBVAR, que utiliza métodos Bayesianos.

Uma boa forma de determinar a ordem de um modelo VAR, ou por outras palavras o atraso p, é observando a autocorrelação da série e a correlação cruzada [27]. Estas medidas permitem-nos avaliar a relevância que uma dada variável independente tem na previsão da nossa variável dependente. A autocorrelação, apresentada na equação 4.2, é uma medida da capacidade de previsão da série em t, utilizando um dado valor passado da mesma série, $t-j$. Se for possível prever perfeitamente Y_t através de um modelo linear dependente apenas de Y_{t-j}, então a autocorrelação entre estas variáveis é 1. A correlação cruzada, descrita em 4.3, é um conceito idêntico ao da autocorrelação, mas entre duas séries temporais, neste caso Y_t e X_t. Este segundo conceito é especialmente relevante em séries temporais multivariadas.

$$\text{autocorrelação}_j = \frac{cov(Y_t, Y_{t-j})}{\sqrt{var(Y_t)var(Y_{t-j})}} \tag{4.2}$$

$$\text{correlação cruzada}_j = \frac{cov(Y_t, X_{t-j})}{\sqrt{var(Y_t)var(X_{t-j})}} \tag{4.3}$$

Para além dos modelos de auto-regressão, são também muito utilizados em análise de séries temporais os modelos MA (*Moving Average*), ou VMA para séries temporais multivariadas, que consistem numa regressão linear do valor actual da série em função de termos actuais e passados de erros que, tipicamente, seguem uma distribuição normal de média nula. O modelo MA de ordem q pode ser descrito da seguinte forma [45]:

$$x_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + ... + \theta_q \varepsilon_{t-q} \tag{4.4}$$

A constante μ corresponde à média da série, os termos em ε são os erros e os termos em θ os coeficientes da equação. Os Modelos VARMA (*Vector Autoregressive Moving-Average*) são uma combinação dos modelo VAR e VMA.

Apesar de os modelos referidos anteriormente serem os mais populares para séries temporais, optou-se por utilizar também redes neuronais e *random forests* por serem modelos não lineares que, apesar de não estarem preparados directamente para séries temporais, demonstraram bons resultados em tentativas passadas de solucionar o problema da Gestão de Descarregamentos [21, 22].

As redes neuronais artificiais são modelos matemáticos baseados conceptualmente nas redes neuronais dos seres vivos. São compostas por neurónios e ligações entre estes, sendo que cada ligação tem um peso associado que permite ao neurónio saber a relevância de cada sinal de entrada. Pode-se dizer que cada neurónio mapeia as suas entradas numa saída fazendo passar a soma pesada

das entradas por uma função de transferência, geralmente uma função logística [46]. A figura 4.1 apresenta o esquema de um neurónio com as entradas pesadas e a função de transferência.

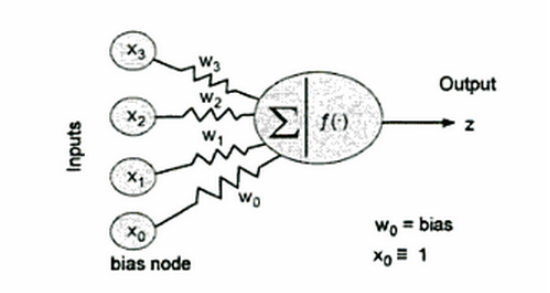


Figura 4.1: Modelo de um neurónio [46]

As *random forests* são um dos métodos mais utilizados de *ensemble learning*, baseadas em árvores de decisão e utilizadas tanto para regressão como para classificação [47]. São uma extensão do método de *bagging*, proposto por [48], que consiste na agregação de vários modelos, cada um criado a partir de um conjuntos de dados seleccionado aleatoriamente e com repetições do conjunto inicial. No caso das *random forests*, os modelos originados são árvores de decisão e em regressão o resultado da previsão é dado pela média dos resultados obtidos em todas as árvores.

Em R, foram utilizados os modelos VAR, VMA e VARMA, do *package* MTS (*Multivariate Time Series*) [49], que inclui as rotinas para modelação e previsão de cada um dos modelos, o modelo SZBVAR disponível no *package* MSBVAR (*Markov-Switching, Bayesian, Vector Autoregression Models*) [50] e o DYNLM do *package* dyn (*dynamic*) [51]. Ao utilizar as funções de previsão do *package* MTS observou-se que estas somente permitem obter previsões a partir de dados utilizados na criação do modelo, variando apenas a origem das previsões de entre este conjunto de dados. As únicas instâncias não observadas que se pode prever com esta função são obtidas escolhendo como origem das previsões o fim do conjunto de dados de modelação. Desta forma, apenas se obteria um conjunto de previsões por modelo para $t+1$ até $t+h$, sendo que t corresponde ao instante da última observação utilizada na modelação. Este comportamento não é desejável para o estudo do desempenho dos modelos obtidos pois a medida de performance deve ser o menos parcial possível. Para que tal aconteça, devem ser feitas várias previsões com dados ainda não observados, calculado o erro de cada uma das previsões e feita a média desses erros. Assim, optou-se por elaborar o código de previsão os para modelos lineares de séries temporais, baseado no código da função VARpred do *package* "MTS", alterando o uso dos dados antigos por novas observações.

Utilizou-se também as redes neuronais do *package* nnet (*neural network*) [52] e as *random forests* do *package* randomForest [53]. Mais à frente no processo de desenvolvimento do trabalho inclui-se também o modelo cforest (*conditional inference random forests*) do *package* party [54, 55, 56]. Os motivos desta inclusão serão apresentados na secção 4.6 mas, em prol de uma comparação abrangente e cuidada, este modelo encontra-se incluído na análise inicial. Para estes algoritmos, dado não estarem adaptados a séries temporais, foi necessário integrar as variáveis atrasadas

como novas variáveis independentes, da mesma forma que [22, 19] fazem nos seus trabalhos.

Com o intuito de comparar cada um dos modelos anteriormente referidos e averiguar a utilidade das previsões, determinou-se também o desempenho de um modelo *naive* cuja previsão dos futuros h valores é igual ao último valor observado. As medidas de performance utilizadas para determinar esse desempenho foram o RMSE e o MAPE, que podem ser determinadas pelas equações 4.5 [15] e 4.6 [57], respectivamente.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2} \tag{4.5}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \widehat{y_i}}{y_i}\right| \tag{4.6}$$

Foi definido pela EDP que o erro máximo aceitável nas previsões em cheia seria de 5% (MAPE) na primeira hora, com menor aumento possível nas restantes, daí o uso desta medida. No entanto, para fazer avaliações em todos os pontos não é possível utilizar o MAPE dado que a caudal afluente é muitas vezes zero ou próximo desse valor. Assim, optou-se por utilizar o RMS para avaliar o comportamento geral, apesar de este não ser o foco do trabalho mas sim as cheias.

Na fase de modelação, os modelos foram criados com dados relativos aos anos de 2001 e 2002 e avaliados com o ano de 2003. Visto o objectivo final ser as previsões em cheias, 2001 é um ano muito importante para modelação visto ter tido uma cheia excecional. O objectivo foi minimizar o RMSE de cada modelo. Para a fase de teste escolheu-se o ano 2010 por ser o único ano depois de 2001 que também contém pontos de cheias, apesar de em quantidade muito mais reduzida. Mais especificamente, enquanto que 2001 possui 133 pontos de cheia, 2010 possui apenas 46. Ainda assim, mesmo em 2001 existe a preocupação de estes pontos não serem suficientes, dado que um ano contém 8760 pontos. Na subsecção 4.2.2 apresenta-se os resultados da fase de modelação e avaliação relativamente ao comportamento global do modelo, ou seja, em cheia e não cheia. Na subsecção 4.2.3 apresenta-se resultados obtidos nas situações de cheias.

4.2 Resultados da 1ª Iteração

Na primeira iteração aplicou-se os algoritmos descritos anteriormente, ao conjunto de dados horários inicial, tendo apenas sido feita a limpeza de dados apresentada na secção 3.4, e utilizando as variáveis seleccionadas na secção 3.5. Os algoritmos de *moving average* foram abandonados pois, após uma quantidade de tempo considerável investida a tentar integrar e resolver os problemas com esses algoritmos, concluiu-se que abandonar esta abordagem seria a melhor opção a tomar. Devido a problemas de implementação das funções em R, seria necessário criar bibliotecas próprias de modelação e previsão, à semelhança do que aconteceu com as funções de previsão dos modelos auto-regressivos mas com complexidade acrescida. Para além disso, devido ao elevado número de variáveis e dimensão dos dados, estes modelos tornam-se demasiado pesados. Ainda assim, a implementação destes modelos fica referenciada como trabalho futuro.

Os resultados obtidos podem ser divididos em globais, sendo estes avaliados com todos os pontos, e em cheias, correspondendo apenas à performance nos pontos de cheia. Optou-se por apresentar o desempenho dos modelos sob a forma de gráficos, onde o eixo das abcissas corresponde ao tempo no qual as previsões são feitas, indo desde uma hora até ao horizonte de previsão, e o eixo das ordenadas à medida de performance indicada. Assim, é possível fazer uma análise comparativa entre os modelos obtidos e o modelo *naive* de uma forma rápida e apelativa. Sempre que necessário, é possível consultar os valores de desempenho obtidos nas tabelas respectivas, presentes no Anexo B. Antes de passar à análise de resultados, apresenta-se uma explicação e conclusões relativas à determinação do atraso, p.

4.2.1 Determinação de p

A primeira tarefa que se abordou foi a determinação, de forma aproximada, de quanto tempo é que a água demora desde o instante em que sai da barragem a montante até chegar à barragem a jusante. Para os nossos modelos, este valor corresponde ao *lag* mínimo a aplicar aos dados mas, para simplificação, vamos chamar a este tempo $t_{percurso}$. Para determinar este factor, elaborou-se um gráfico de correlação cruzada entre a água que é libertada pela barragem a montante (caudal descarregado mais caudal turbinado) e o caudal afluente da barragem a jusante. Em primeiro lugar, esta correlação deve ser positiva pois quanto mais água é libertada a montante mais água chega a jusante, o que implica que o caudal afluente é também ele maior. Para além disso, se a correlação cruzada for calculada relativamente aos atrasos do caudal libertado a montante face ao afluente em jusante no instante zero, prevê-se que haja um pico de correlação num instante de tempo negativo e que este vá diminuindo mais ou menos simétricamente. Esse pico de correlação deverá ocorrer para o atraso correspondente ao $t_{percurso}$. Por outras palavras, a água que chega à barragem mais a jusante num dado instante é fortemente relacionada com a água libertada pela barragem mais a montante à $t_{percurso}$ horas atrás.

Na imagem 4.2 observa-se que o formato do gráfico é idêntico ao esperado mas que o pico se dá em zero e não num valor negativo, como previsto. Isto implica que o valor que mais influencia a média do caudal afluente de uma dada hora na barragem mais a jusante é a média do caudal libertado pela barragem a montante nessa mesma hora. Para explicar este fenómeno colocou-se a hipótese de as barragens estarem a uma distância curta o suficiente para que o $t_{percurso}$ seja de uma hora ou menos. Esta hipótese foi confirmada pelos peritos no domínio e leva a uma conclusão que pode influenciar os resultados de todo o trabalho: a granularidade dos dados poderá ser demasiado grossa para se obter boas previsões. Intuitivamente é fácil de inferir que, para relacionar a água que sai de uma dada barragem com a que chega à barragem seguinte, é necessário que os dados tenham sido amostrados com um período inferior ao $t_{percurso}$. Segundo o Teorema de Nyquist, a frequência de amostragem de um sinal contínuo deve ser pelo menos o dobro da frequência mais alta desse sinal [58]. Por outras palavras, se considerarmos um processo real e contínuo de período T então o tempo decorrido entre duas amostras consecutivas deste sinal deverá ser no máximo $T/2$. Conceptualmente, pode-se adaptar este teorema ao caso de estudo considerando que o fluxo de água de uma barragem para a outra é o processo real que se quer amostrar. Assim,

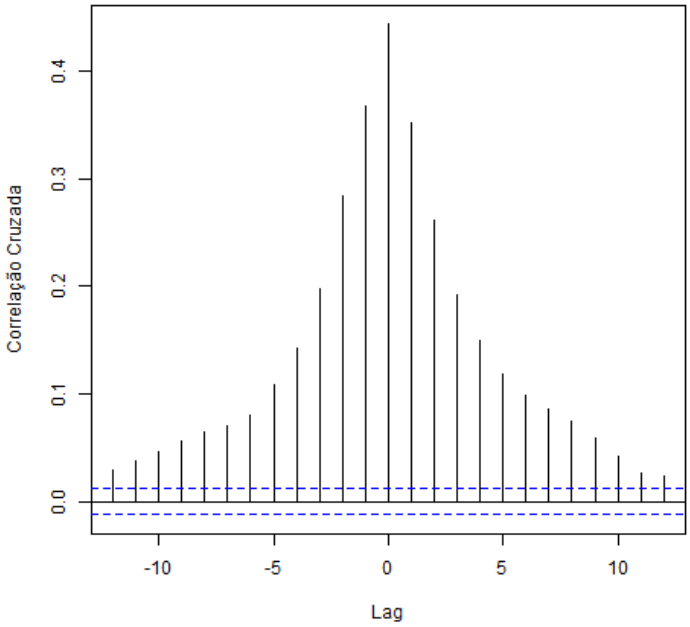


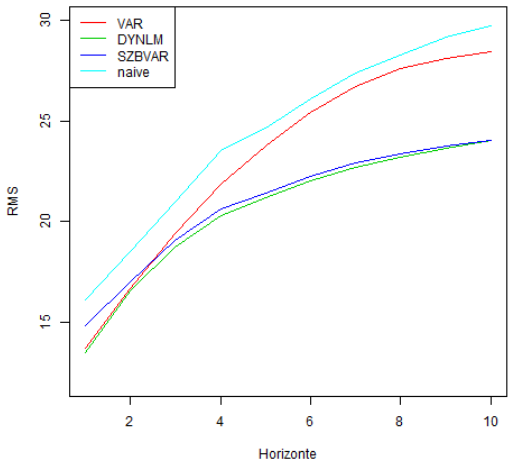
Figura 4.2: Gráfico de correlação cruzada entre o caudal afluente em Venda Nova e o libertado em Alto Rabagão

admitindo que $t_{percurso}$ é de uma hora, podemos dizer que este fenómeno tem o período igual a uma hora e que, portanto, deve ser amostrado no máximo de meia em meia hora.

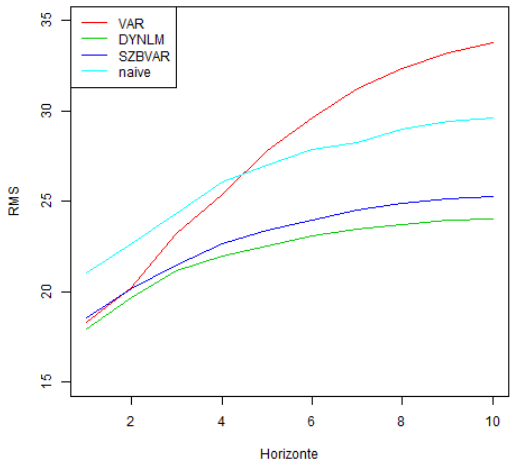
Apesar desta dificuldade, optou-se por continuar com a abordagem planeada inicialmente visto os restantes valores de correlação cruzada também apresentarem uma magnitude significativa o suficiente para indicar que poderá ser possível fazer previsões com base apenas nos valores passados. Dado que não foi possível determinar um p mínimo, este parâmetro foi determinado para cada algoritmo em modelação. Em seguida, apresenta-se os resultados finais obtidos nesta fase e em avaliação. Para além disso, na subsecção 4.2.4 faz-se uma comparação entre o desempenho obtido com e sem recursividade nas previsões.

4.2.2 Resultados globais

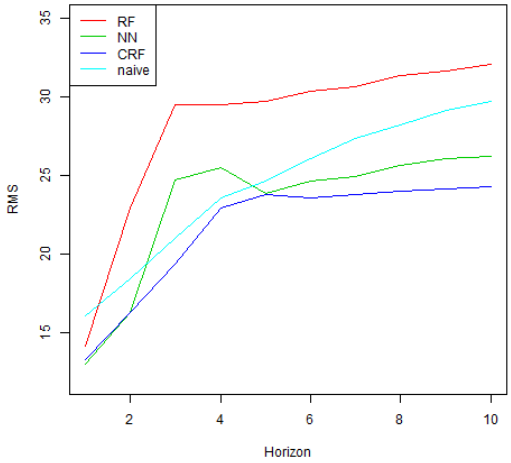
Nesta secção apresenta-se os resultados globais obtidos na primeira iteração, tanto em modelação como em avaliação. Numa primeira fase pode ser interessante comparar o desempenho obtido nestas duas fases, apesar de, para os restantes casos, ser mais simples apresentar apenas o resultado em avaliação, dado este ser o que mede o desempenho do modelo com dados ainda não observados. Para além disso, não é possível avaliar o desempenho em cheia no caso da modelação dado estar-se a usar os dados de 2001 e 2002 para fazer os modelos e os de 2003, que não têm cheias, para a avaliação. A performance obtida tanto com os modelos lineares como com os modelos não lineares nestas duas etapas pode ser observada na figura 4.3.



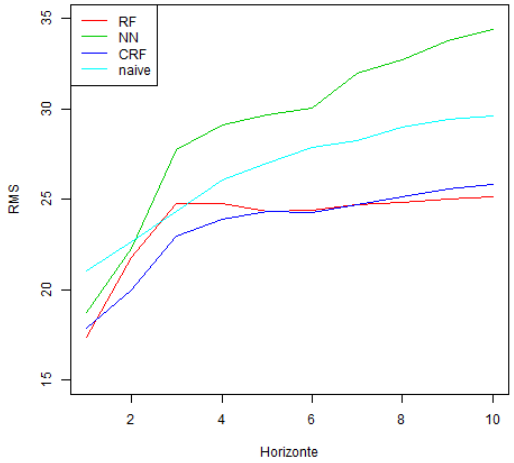
(a) RMSE em modelação dos modelos lineares



(b) RMSE em avaliação dos modelos lineares



(c) RMSE em modelação dos modelos não lineares



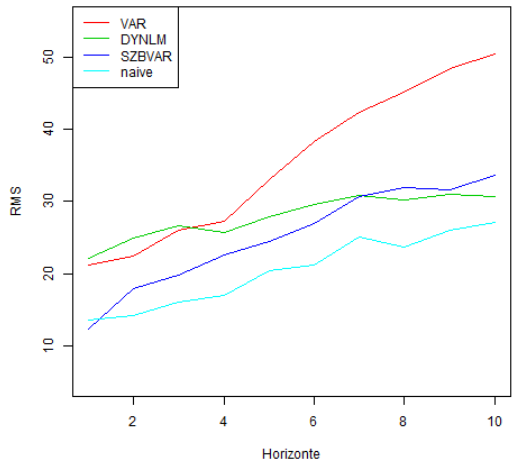
(d) RMSE em avaliação dos modelos não lineares

Figura 4.3: RMSE obtido com os algoritmos séries temporais para todos os pontos

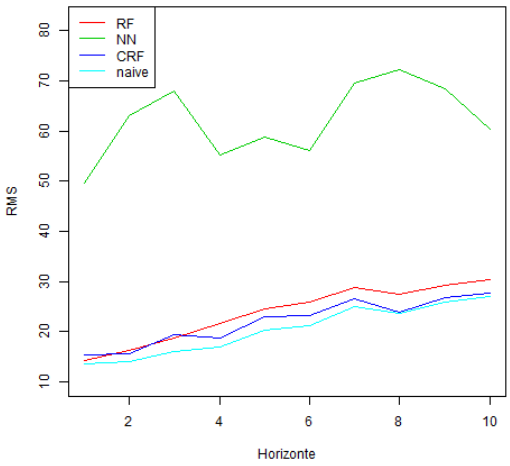
Comparando as figuras 4.3a e 4.3b podemos concluir que apenas o modelo VAR se comporta significativamente pior na fase de avaliação do que na de modelação, sendo este o único que apresenta pior desempenho que o modelo *naive*. Relativamente aos algoritmos não lineares, comparando as figuras 4.3c e 4.3d, pode-se concluir que as redes neuronais são as que apresentam pior desempenho em ambas as fases, tendo pior performance em avaliação que em modelação, como esperado. Pelo contrário os modelos RF e CRF comportam-se melhor em avaliação do que em modelação, contra o que seria de esperar, apresentando um comportamento bastante aceitável em avaliação. Este fenómeno poderá ser devido ao facto da avaliação refletir o comportamento global e, dado que em modelação o modelo é criado com um ano de cheia excecional e um sem, este ser mais afetado do que em avaliação, em que se utiliza mais um ano normal do que anteri-

ormente. Portanto, para retirar conclusões úteis para o problema em questão é necessário medir o desempenho destes modelos em cheia.

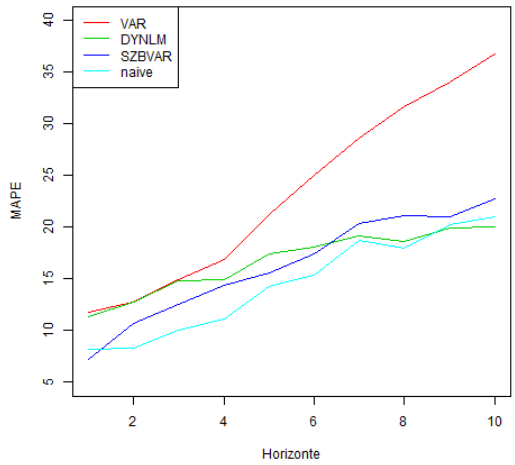
4.2.3 Resultados em cheia



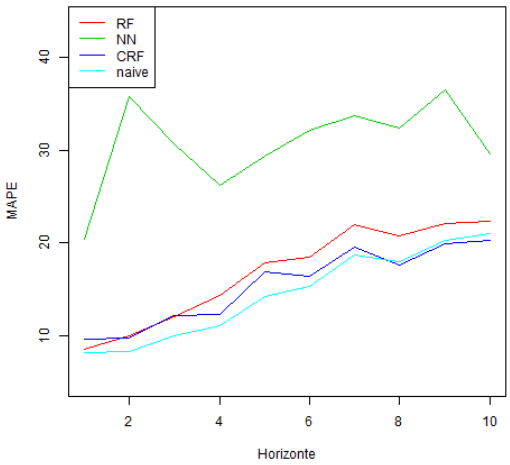
(a) RMSE modelos lineares



(b) RMSE modelos não lineares



(c) MAPE modelos lineares



(d) MAPE modelos não lineares

Figura 4.4: Performance em avaliação para os algoritmos de séries temporais, nos casos de cheia

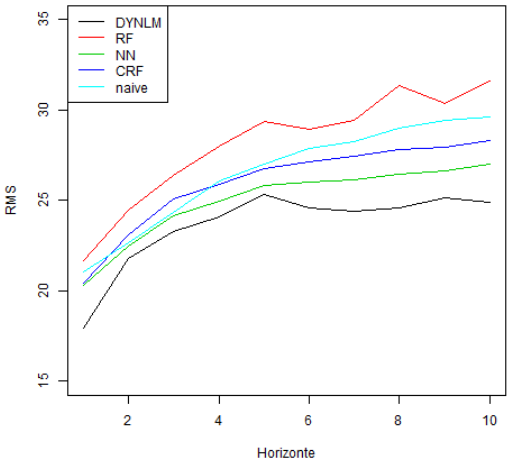
O conjunto de figuras 4.4 apresenta os resultados da fase de avaliação apenas nos pontos de cheia. Destes gráficos conclui-se que, apesar de os modelos lineares e alguns dos modelos não lineares apresentarem um desempenho aceitável no caso geral, em cheia o mesmo já não se passa, sendo raro estes apresentarem resultados melhores que o modelo *naive*. Este resultado leva a crer que as cheias não estão a ser bem modeladas pois os modelos focam-se em maximizar o

desempenho geral. Este efeito faz-se sentir especialmente no caso das redes neuronais. Isto não só vai de encontro com o esperado como valida o uso de métodos específicos para tratar as cheias como a regressão de quantis e a previsão de valores raros extremos, apresentadas na secções 4.5 e 4.6, respectivamente.

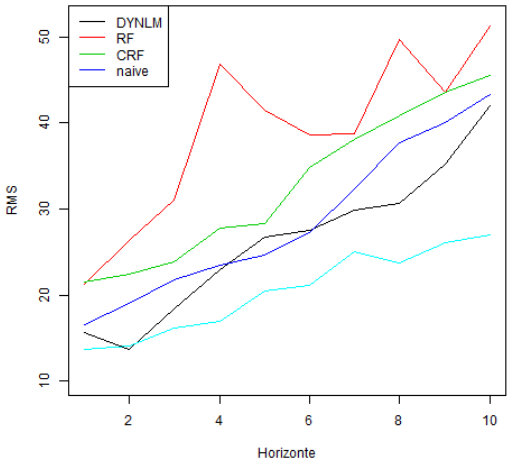
4.2.4 Resultados sem recursividade

Na subsecção 2.3.2.1 explicou-se o funcionamento geral das séries temporais, referindo-se que existem duas possibilidades para obter previsões a mais de um passo. Os resultados apresentados até agora utilizam as previsões realizadas para substituir os valores observados nos instantes passados, fazendo assim com que o erro seja cumulativo e vá aumentando à medida que se aumenta o horizonte de previsão, podendo estabilizar para valores altos do horizonte. Para comprovar que esta é a melhor opção para o nosso estudo, optou-se por repetir as experiências anteriores utilizando o método referido na equação 2.4. Assim, em vez de um modelo para cada variável, vamos ter tantos modelos quantos o horizonte de previsão exigir, mas apenas para a variável objetivo, que neste caso é o caudal afluente. Esta variante não foi aplicada aos modelos específicos de séries temporais, como o VAR, dado que as suas funções não o permitem. Como nos restantes modelos, as variáveis atrasadas não estão integradas mas são adicionadas como qualquer outra variável independente, é possível utilizar esta abordagem. A figura 4.5 apresenta os resultados obtidos e a tabela 4.1 contém a diferença entre a performance resultante deste método e a obtida anteriormente.

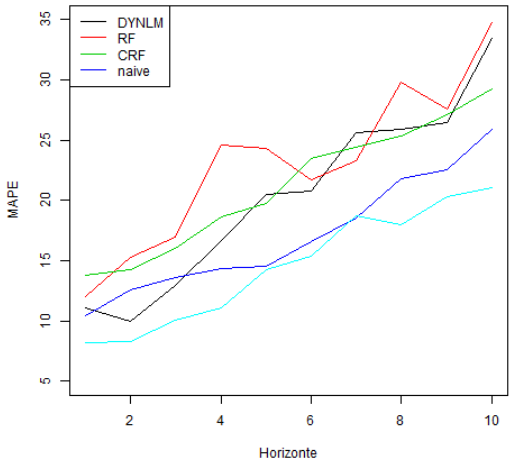
No caso geral o MAPE aumentou, significando que o desempenho piorou, exceto para as redes neuronais. Possivelmente, existem algumas variáveis que as redes neuronais têm mais dificuldade em prever do que o afluente e, dado que desta forma apenas se prevê o caudal afluente, esse problema foi corrigido. Em cheia todos pioraram bastante, novamente com a excepção das redes neuronais. Em relação ao horizonte considerado, a maioria dos modelos piora a longo prazo, apesar de a diferença não ser significativa a curto prazo. No entanto, em termos práticos, é muito mais vantajoso criar um modelo para cada variável e utilizar o método recursivo dado que desta forma é possível aumentar o horizonte de previsão sem alterar os modelos, sendo por isso mais versátil. Neste caso em que o horizonte é 10 e o número de variáveis também, em qualquer dos métodos o nº de modelos é o mesmo. Mas, por exemplo, se por algum motivo for preciso alterar o horizonte de previsão para 48 horas (2 dias), enquanto que com a abordagem recursiva isso seria feito apenas aumentando o número de previsões, com a abordagem não recursiva seria necessário aumentar o número de modelos para 48. Em suma, tanto a nível de desempenho como em versatilidade, é mais vantajoso utilizar o método recursivo.



(a) RMSE geral



(b) RMSE em cheia



(c) MAPE em cheia

Figura 4.5: Performance em avaliação sem recursividade

Tabela 4.1: Diferença entre o MAPE obtido com e sem recursividade

Modelos	Horizonte									
	1	2	3	4	5	6	7	8	9	10
DYNLM	-0.29	-2.84	-1.87	1.81	3.07	2.71	6.48	7.31	6.51	13.40
RF	3.51	5.30	4.94	10.26	6.43	3.19	1.38	8.99	5.40	12.41
NN	-6.71	-21.57	-14.67	-7.59	-9.54	-8.78	-9.33	-7.01	-9.37	-0.42
CRF	0.85	2.91	1.42	2.06	-2.36	0.13	-1.07	4.23	2.57	5.55

4.3 Séries Temporais com novas variáveis

Antes de passar para novas abordagens optou-se por tentar fazer adaptações aos dados disponíveis, utilizando os mesmos algoritmos que anteriormente. Em seguida, apresenta-se essas adaptações e os seus resultados.

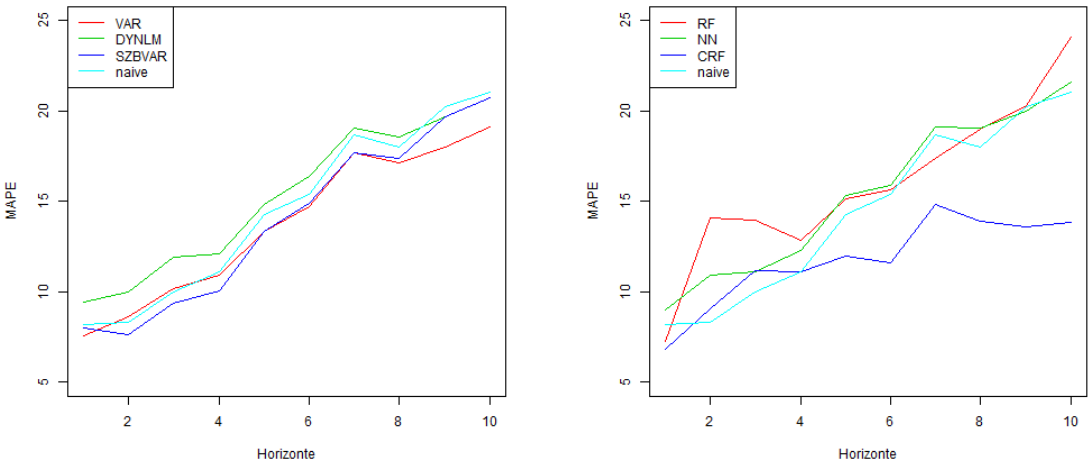
4.3.1 Variações

Em alguns casos cuja variável dependente é contínua, obtém-se melhores resultados modelando as variações desta do que o seu valor absoluto. Assim, decidiu-se aplicar os mesmos modelos descritos anteriormente ao conjunto de dados derivado do original, em que se substitui cada instância de cada variável pela diferença entre o seu valor e o valor anterior. Assim, passa-se a representar não o valor absoluto naquela hora, mas sim a sua variação relativamente à hora anterior e o objetivo passa a ser prever a variação na próxima hora do caudal afluente. De forma a obter resultados comparáveis com os anteriores, as previsões são somadas ao último valor observado do caudal afluente, obtendo novamente o módulo. Nas previsões seguintes, utiliza-se a última previsão em vez do último valor observado, como representado na equação 4.7.

$$\widehat{x_t} = \widehat{x}_{t-1} + \Delta\widehat{x_t} \tag{4.7}$$

em que $\widehat{x}$ correspondem as previsões do caudal afluente e $\Delta\widehat{x}$ às previsões da variação do mesmo.

Na figura 4.6 apresenta-se os resultados obtidos em avaliação apenas nos casos de cheia. Optou-se por não apresentar aqui novamente a performance no caso geral, dado o objectivo ser melhorar o desempenho em cheia. No entanto, é possível consultar todos os resultados obtidos no Anexo A. A tabela 4.2 contém a diferença entre os resultados aqui obtidos e os da secção 4.2.3, de forma a facilitar a comparação entre estes. Analisando estes dois elementos, pode-se concluir que foi possível obter melhor desempenho utilizando as variações. Como explicado anteriormente, esta melhoria deve-se ao facto de os algoritmos terem mais facilidade em modelar as variações pois estas possuem uma gama mais reduzida de valores e incorporam directamente a informação da tendência. Para além disso, as melhorias são mais significativas em cheias dado que é quando ocorrerem maiores variações. Este efeito é especialmente notório nas redes neuronais que chegam a ter mais de 40% de melhoria no desempenho nas últimas previsões. Ainda assim, apenas três dos modelos obtidos apresentam um desempenho superior ou semelhante ao modelo *naive*, sendo estes o modelo, VAR, SZBVAR e CRF.



(a) Modelos lineares

(b) Modelos não lineares

Figura 4.6: MAPE em avaliação nos casos de cheia, com as variáveis correspondentes às variações

Tabela 4.2: Diferença entre o MAPE obtido com e sem as variações

Modelos	Horizonte									
	1	2	3	4	5	6	7	8	9	10
VAR	-4.20	-4.05	-4.79	-5.90	-7.91	-10.29	-10.95	-14.54	-16.08	-17.69
DYNLM	-1.97	-2.77	-2.88	-2.76	-2.64	-1.69	-0.07	-0.06	-0.26	0.71
BVAR	0.81	-3.09	-3.19	-4.36	-2.17	-2.54	-2.63	-3.77	-1.38	-1.97
RF	-1.28	4.14	1.94	-1.48	-2.74	-2.83	-4.54	-1.82	-1.86	1.74
NN	-14.12	-21.55	-31.30	-26.92	-30.73	-30.69	-32.28	-39.86	-42.96	-46.08
CRF	-2.79	-0.65	-1.08	-1.19	-4.89	-4.79	-4.75	-3.69	-6.37	-6.46

4.3.2 Variáveis esporádicas

Para além dos dados horários, foi também fornecido um conjunto de dados esporádicos, com registos do valor instantâneo de algumas variáveis, em momentos não pré-determinados. Apesar de não ter uma base tempo fixa, sabe-se que o valor de uma dada variável é registado quando esta tem uma variação, relativamente ao valor anterior, superior a um dado limiar. Assim, apesar de não ser fácil de integrar diretamente esta informação no conjuntos de dados horários, é possível extrair outras informações potencialmente benéficas, nomeadamente o último valor registado para cada hora. Dado os valores horários corresponderem à média de cada variável na última hora, espera-se que o seu valor numa dada hora seja mais próximo do último valor registado da hora anterior do que da sua média, nos casos em que existem variações. Para criar este novo conjunto de dados utilizou-se o MATLAB. A tabela 4.3 apresenta um exemplo da conversão feita dos dados esporádicos para o último valor horário registado. Esta conversão foi realizada apenas para as variáveis registadas nos dados esporádicos, sendo estes o caudal descarregado e turbinado de ambas as barragens.

Tabela 4.3: Exemplo de conversão dos dados esporádicos para último valor registado horário

(a) Extracto dos dados esporádicos de Alto Rabagão

Data (ano.mês.dia)	Hora (hora.minutos)	Caudal Descarregado
01.03.22	7.39	126
01.03.22	7.43	109
01.03.22	9.01	51
01.03.23	15.58	83
01.03.24	9.5	51
01.03.26	10.1	0

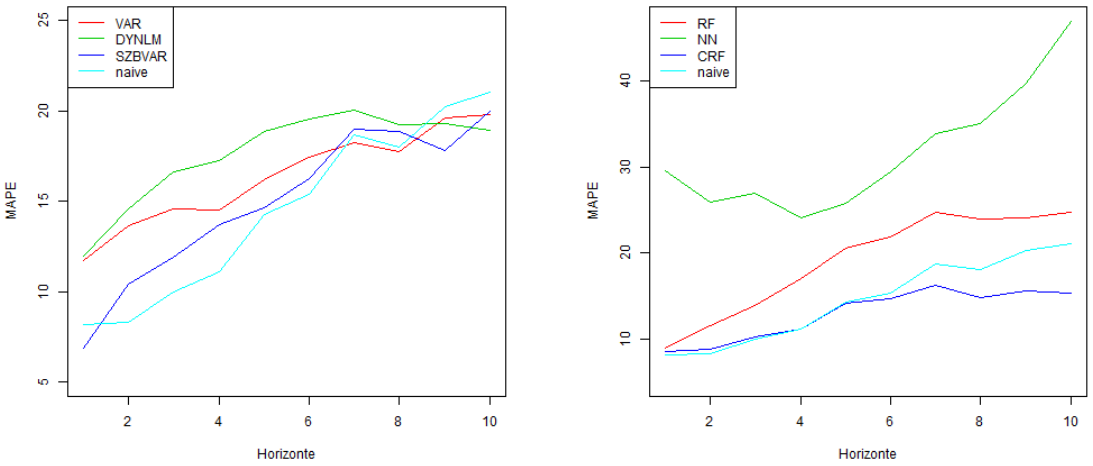
(b) Dados horários após conversão

Data (ano.mês.dia)	Intervalo horário	Último Caudal Descarregado
01.03.22	8 - 9	109
01.03.22	10 - 23	51
01.03.23	0 - 15	51
01.03.23	16 - 23	83
01.03.24	0 - 9	83
01.03.24	10 - 23	51
01.03.25	0 - 23	51
01.03.26	0 - 10	51
01.03.26	11	0

Na figura 4.7 pode-se observar os resultados obtidos acrescentando estas quatro variáveis aos modelos anteriores e na tabela 4.4 a diferença entre o MAPE resultante desta experiência e o obtido na subsecção 4.2.3. Da sua análise conclui-se que, apesar de haver uma melhoria significativa nas últimas previsões do modelo VAR e NN, as restantes alterações na performance não são suficientemente significativas para afirmar que esta experiência realmente melhora os resultados da previsão. Para além disso, de uma forma geral os modelos comportam-se pior que o modelo *naive*. Apenas o modelo CRF mostra uma melhoria ao longo do horizonte e é o único que apresenta resultados positivos face ao modelo *naive*.

Tabela 4.4: Diferença entre o MAPE obtido com e sem o último valor registado das variáveis esporádicas

Modelos	Horizonte									
	1	2	3	4	5	6	7	8	9	10
VAR	0.00	0.94	-0.35	-2.28	-5.07	-7.55	-10.39	-13.91	-14.47	-16.98
DYN	0.63	1.81	1.85	2.41	1.39	1.49	0.92	0.62	-0.66	-1.11
BVAR	-0.31	-0.26	-0.62	-0.69	-0.89	-1.18	-1.35	-2.26	-3.19	-2.72
RF	0.40	1.64	1.82	2.72	2.65	3.41	2.81	3.14	1.99	2.37
NN	6.49	-6.53	-15.51	-15.04	-20.32	-17.22	-17.46	-23.91	-23.25	-20.70
CRF	-1.00	-0.87	-1.96	-1.18	-2.70	-1.70	-3.38	-2.82	-4.34	-5.03



(a) Modelos lineares

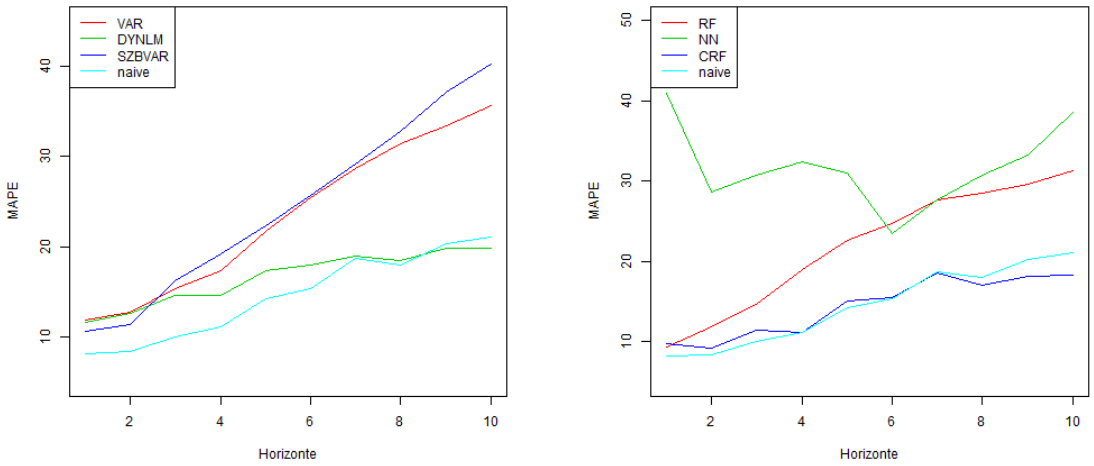
(b) Modelos não lineares

Figura 4.7: MAPE em avaliação nos casos de cheia, com as variáveis do último valor registado

Para além do último valor registado numa dada hora, foi também criado um novo conjunto de variáveis horárias relativas ao caudal descarregado e turbinado, que representam a sua variância dentro de uma dada hora. Nas horas em que as variáveis no conjunto de dados esporádicos não têm ocorrências, considerou-se a variância como sendo zero. Nos restantes casos, calculou-se a variância entre todos os registos daquela hora e o último registo da hora anterior, ou ocorrência mais próxima caso a hora anterior não tenha registos. A variância foi determinada utilizando o comando "var"do MATLAB [59] que, para um dado vector A com N observações e média μ, realiza o seguinte cálculo:

$$V = \frac{1}{N-1} \sum_{i=1}^{N} |A_i - \mu|^2 \tag{4.8}$$

A figura 4.8 apresenta os resultados obtidos acrescentando estas quatro variáveis aos modelos anteriores e a tabela 4.5 a diferença entre o MAPE obtido com a inclusão destas variáveis face ao que foi obtido na subsecção 4.2.3, sem estas.



(a) Modelos lineares

(b) Modelos não lineares

Figura 4.8: MAPE em avaliação nos casos de cheia, com a variância

Tabela 4.5: Diferença entre o MAPE obtido com e sem a variância das variáveis esporádicas

Modelos	Horizonte									
	1	2	3	4	5	6	7	8	9	10
VAR	0.12	0.05	0.38	0.45	0.45	0.37	0.02	-0.33	-0.72	-1.22
DYN	0.21	-0.21	-0.23	-0.21	-0.18	-0.16	-0.16	-0.21	-0.16	-0.19
BVAR	3.44	0.68	3.63	4.75	6.74	8.25	8.84	11.58	16.05	17.49
RF	0.89	1.92	2.64	4.63	4.70	6.20	5.72	7.77	7.43	8.91
NN	17.93	-3.79	-11.60	-6.79	-15.05	-23.07	-23.64	-28.22	-29.73	-29.06
CRF	0.13	-0.42	-0.79	-1.09	-1.76	-0.87	-1.00	-0.60	-1.76	-1.97

Analisando a figura 4.8 e a tabela 4.5 pode-se concluir que acrescentar estas variáveis praticamente não altera o desempenho de alguns modelo, como é o caso dos modelos VAR, DYNLM e CRF, ou então prejudica-o significativamente, como fez nos modelos restantes. Apesar de teoricamente todos estes algoritmos serem suficientemente insensíveis a variáveis irrelevantes [60, 61] e, sendo assim, seria expectável que o seu desempenho não piorasse significativamente pelo acrescentar de novas variáveis, pode-se concluir que os modelos que foram mais afectados são algo sensíveis a este tipo de variáveis. Por exemplo, os modelos de regressão ou de redes neuronais poderão estar a dar pesos baixos às variáveis irrelevantes mas não o suficiente para que estas deixem de influenciar o modelo [62]. Para além disso, a variância é na grande maioria dos casos zero, o que dificulta a sua modelação e previsão, podendo requerer algoritmos próprios como os modelos *zero-inflated* que são direcionados para a modelação de dados com elevada frequência de zeros [63].

4.4 Granularidade mais fina

Em resposta ao problema da granularidade encontrado na subsecção 4.2.1, decidiu-se criar um conjunto de dados com granularidade mais fina, a partir dos dados horários e esporádicos, sendo a periodicidade escolhida de 15 minutos. Para as variáveis que não se encontram registadas no conjunto de dados esporádicos, optou-se por igualar os quatro novos registos de uma dada hora à sua média horária, considerando assim que durante essa hora o valor da variável se manteve constante sendo, por isso, a média igual em qualquer instante do tempo. Foi necessário ter em conta que a média horária em t corresponde à média de valores registados durante $t-1$, para que os registos horários e esporádicos ficassem temporalmente concordantes. No caso das variáveis disponíveis no conjunto de dados esporádicos procedeu-se de forma semelhante ao que foi feito para o registo do último valor de caudal descarregado, na subsecção 4.3.2, tendo em atenção que, em vez se ter em conta apenas a hora, é preciso analisar cada intervalo de quarto de hora. Para arredondar os minutos optou-se pelos seguintes intervalos: $[0;14] = 0$, $[15;29] = 15$, $[30;44] = 30$ e $[45;59] = 45$. A Tabela 4.6 apresenta um exemplo da conversão feita dos dados esporádicos para o último valor horário registado.

Tabela 4.6: Exemplo de conversão dos dados esporádicos em dados de 15 em 15 minutos

(a) Extracto dos dados esporádicos de Alto Rabagão

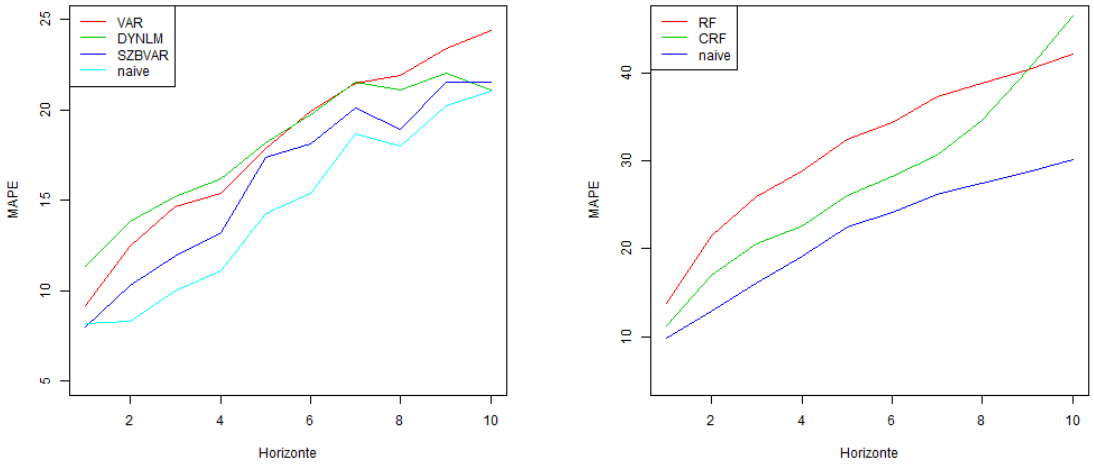
Data (ano.mês.dia)	Hora (hora.minutos)	Caudal Descarregado
01.03.21	12.19	74
01.03.21	12.59	109
01.03.21	13.24	176
01.03.21	13.56	176

(b) Dados com periodicidade de 15 minutos após conversão

Data (ano.mês.dia)	Hora (hora.minutos)	Caudal Descarregado
01.03.21	12.15	74
01.03.21	12.30	74
01.03.21	12.45	109
01.03.21	13.00	109
01.03.21	13.15	176
01.03.21	13.30	176
01.03.21	13.45	176

Assim, as previsões passam a ser feitas de 15 em 15 minutos para todas as variáveis. No entanto, como a avaliação tem de ser feita face a valores observados, para cada conjunto de quatro previsões do caudal afluente, realizou-se a sua média e comparou-se com os dados horários reais. Os resultados obtidos com esta abordagem podem ser observados na figura 4.9 e a sua comparação com os resultados originais na tabela 4.7. A partir da análise destes dois elementos, pode-se concluir que existe uma melhoria nos resultados de alguns modelos, como o modelo VAR

e as redes neuronais, sendo que os outros mantêm ou até mesmo pioraram o seu desempenho em cheias. Apenas analisando estes resultados, não é possível concluir com certeza que esta experiência melhore ou não os resultados. No entanto, dado o novo conjunto de dados ter sido derivado de outros dados e não recolhido, este acarreta erros inerentes às aproximações feitas. Em primeiro lugar, como referido anteriormente, foi apenas possível passar o caudal descarregado e turbinado de ambas as barragens para a granularidade mais fina, dado estas serem as variáveis disponíveis no conjunto de dados esporádicos. Consequentemente, as restantes variáveis incluindo o caudal afluente, variável que se está a avaliar, foram simplesmente replicadas quatro vezes em cada hora. Isto não só provoca um erro significativo nestas variáveis como aumenta a complexidade da modelação, dado que os valores são iguais quatro a quatro. Ainda assim, com todas estas aproximações cometidas na criação do conjunto de dados e consequente modelação, observa-se que na avaliação os modelos não foram muito prejudicados, sendo por vezes beneficiados. Pode-se assim especular que, se fosse possível aproximar todos os dados à frequência de 15 minutos, ou ainda melhor, obter dados reais com essa periodicidade, os resultados iriam melhorar significativamente.



(a) Modelos lineares

(b) Modelos não lineares

Figura 4.9: MAPE em avaliação nos casos de cheia, com a granularidade mais fina

Tabela 4.7: Diferença entre o MAPE obtido com os dados de 15 em 15 minutos e com os dados horários

Modelos	Horizonte									
	1	2	3	4	5	6	7	8	9	10
VAR	-2.60	-0.23	-0.27	-1.41	-3.38	-5.11	-7.20	-9.71	-10.71	-12.40
DYNLM	0.00	1.08	0.45	1.32	0.74	1.68	2.45	2.46	2.07	1.10
BVAR	0.83	-0.42	-0.66	-1.18	1.84	0.69	-0.23	-2.19	0.52	-1.20
RF	5.30	11.53	13.95	14.54	14.57	15.84	15.38	17.94	18.14	19.69
NN	-11.85	-15.50	-21.89	-16.61	-20.02	-18.32	-20.71	-24.37	-22.66	-21.30
CRF	0.29	3.21	3.91	6.83	5.67	7.66	6.55	9.78	8.76	9.76

4.5 Regressão de Quantis

Na subsecção 2.3.2.2 explicou-se o funcionamento da regressão dos quantis, nomeadamente o facto de este método ser vantajoso no caso de existir uma gama da variável dependente mais relevante que as outras. Como foi possível observar na subsecção 3.3.1, as cheias abrangem os valores mais altos das variáveis hidrológicas, sendo o seu comportamento médio muito mais elevado do que o comportamento médio geral. Assim, é de esperar que nos casos de cheia o erro seja consistentemente negativo. Por outras palavras, dado que os modelos elaborados com o conjunto completos de dados tendem a ajustar-se ao comportamento médio, em situações de cheia estes irão prever valores mais baixos que a realidade. Aplicando a regressão de quantis para um valor de q alto, como por exemplo $q = 75\%$, o modelo irá ajustar-se não aos valores médios mas sim aos valores do último quartil da distribuição da variável independente. Desta forma, será mais fácil prever valores elevados.

A regressão de quantis foi aplicada a apenas três dos modelos anteriores, dado que nem todos possuem uma implementação deste método. Os algoritmos escolhidos com possibilidade de integrar os quantis são: dynlm [64], redes neuronais [65] e *random forests* [66]. Na figura 4.10 apresenta-se os resultados obtidos com estes algoritmos, para o quartil 75%, e na tabela 4.8 a diferença entre o desempenho aqui obtido e o inicial da subsecção 4.2.3.

Tabela 4.8: Diferença entre o MAPE obtido utilizando quantis e sem estes

Modelos	Horizonte									
	1	2	3	4	5	6	7	8	9	10
DYNLM	4.78	6.44	9.59	16.82	20.11	25.55	31.63	36.49	40.26	44.70
NN	-6.20	-21.39	-13.34	-8.31	-7.18	-8.58	-7.02	-4.99	-7.73	-0.29
RF	9.20	15.63	21.12	27.55	34.93	50.96	79.21	154.91	256.06	305.32

Ao contrário do que seria de esperar, os resultados em vez de melhorarem pioram significativamente, à excepção das redes neuronais que mostram melhorias mas mesmo assim continuam piores que o modelo *naive*. Para além disso, pode-se concluir pela figura 4.11 que o quantil que melhores resultados obtém é o 0.5, que corresponde à mediana. Numa tentativa de averiguar o motivo destes resultados elaborou-se dois gráficos relativos às previsões em cheia de 2010, contendo

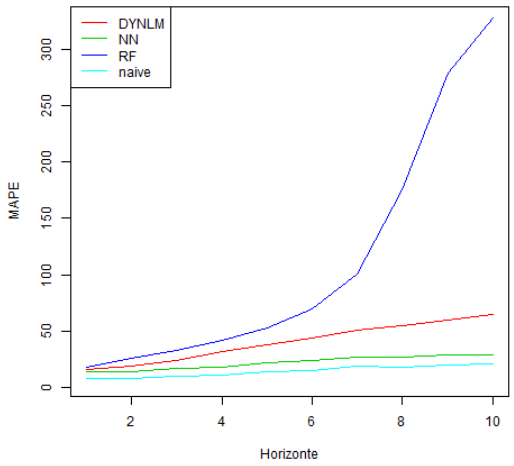


Figura 4.10: MAPE em avaliação nos casos de cheia, com os quantis

os valores reais e os previstos com e sem os quantis. Observando a figura 4.12, que contém esses gráficos, conclui-se que a premissa inicial de os valores em cheia estarem a ser previstos sempre inferiormente está incorreta. Dado que o erro de previsão em 4.12a tanto é positivo como negativo, ao aplicar os quantis o erro absoluto médio irá aumentar pois está-se a fazer com que todas as previsões sejam maiores, como se pode observar em 4.12b pela subida da linha das previsões.

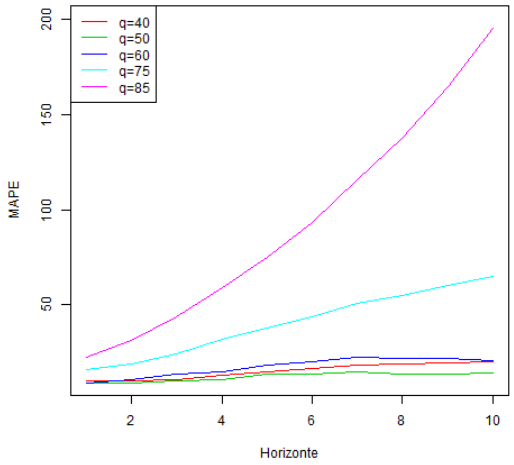
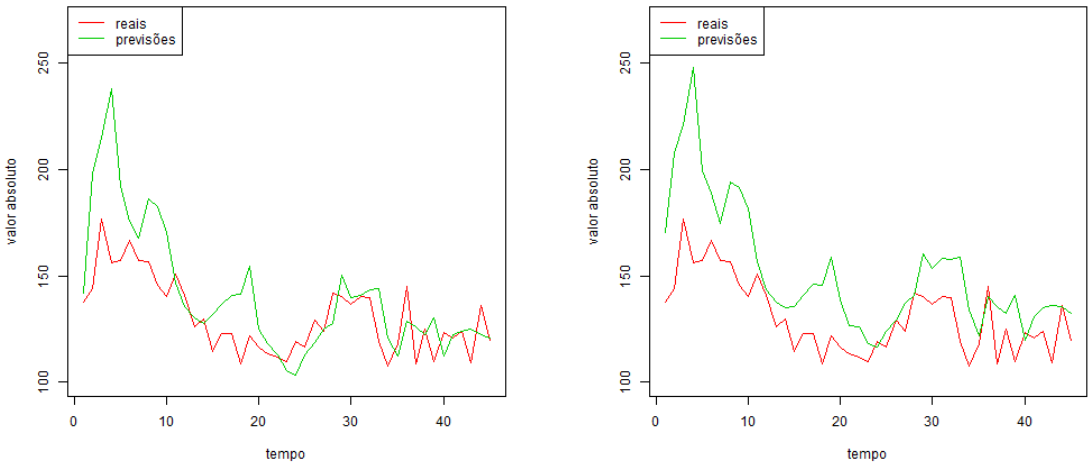


Figura 4.11: Desempenho obtido com o dynlm para vários quantis, nos casos de cheia

Com este resultado chega-se à possibilidade de a distribuição de valores que se obtém de 2001 a 2003 não ser representativa de 2010. Esta conclusão é logicamente válida pois, visto que em 2001 ocorreu uma cheia excecional, os valores superiores desta distribuição irão ser mais altos que os valores das cheias de 2010 que, apesar de serem raras, podem não ser tão elevadas. Novamente se levanta o problema da falta de pontos de cheia, que neste caso é agravada pelo facto de a gama



(a) sem quantis (b) com quantis

Figura 4.12: Previsões a uma hora das cheias com o 3º quartil e valores reais

de valores que se quer priorizar poder não ser semelhante à gama de valores que se está a avaliar. Ou seja, seria preciso duas cheias excecionais ou duas cheias consideradas normais para realizar um modelo e avaliá-lo em condições semelhantes.

4.6 Previsão de valores raros extremos

Foi salientado na subsecção 2.3.2.3 que, ao passo que da regressão de quantis foca apenas em melhorar o desempenho para um dada gama de valores da variável dependente, na previsão de valores raros extremos tem-se em consideração se este é um valor raro ou não. Ou seja, na regressão de quantis privilegiou-se todos os valores altos da variável dependente, quer estes fossem cheias ou não, ao passo que aqui se pretende focar apenas nas cheias. Assim, pode-se forçar os algoritmos a focar os casos de cheias, quer estes apresentem valores altos ou não. Optou-se por tomar uma abordagem simples, que consiste em atribuir maior peso aos casos de interesse. Por defeito, a função de optimização dos modelos minimiza o erro global, dando a mesma importância a todos os pontos. À semelhança do *Cost-Sensitive Learning* para classificação, em regressão pode-se atribuir pesos aos exemplos que consideramos os mais importantes e, neste caso, os mais raros. Assim, a optimização do modelo é feita privilegiando esses casos em detrimento dos outros. Como apontado em [38], atribuir mais peso aos casos raros promove a sua previsão, melhorando a performance nestes pontos, mas não penaliza a previsão de valores extremos quando a realidade é um valor normal. Para a aplicação em causa, visto que o objectivo final é fazer previsões em cheias, esta contrapartida não é prejudicial, podendo este método ser aplicado sem nenhuma restrição.

De entre os modelos estudados inicialmente, apenas o DYNLM e redes neuronais permitiam atribuir pesos aos exemplos. Tendo em conta que o DYNLM é um bom representante dos modelos

lineares, ficava apenas a faltar um exemplo das *random forests*. Assim, decidiu-se acrescentar aos modelos de estudo o CRF, por fazer parte da família das *random forests* e permitir atribuir pesos aos exemplos. A figura 4.13 mostra o desempenho destes modelos, atribuindo às cheias o dobro do peso que aos restantes casos, e a tabela 4.9 apresenta a diferença entre a performance obtida com e sem os pesos.

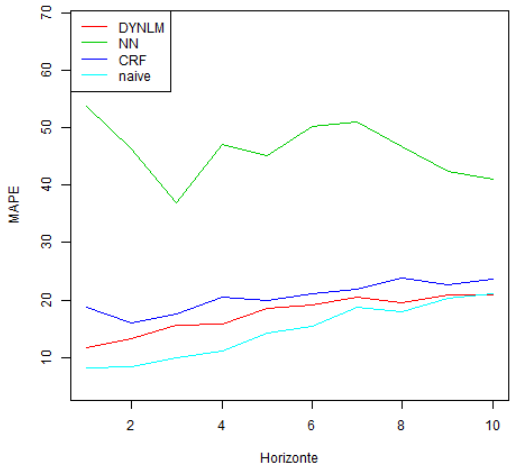


Figura 4.13: MAPE em avaliação nos casos de cheia, com os pesos

Tabela 4.9: Diferença entre o MAPE obtido aplicando pesos às cheias e sem estes

Modelos	Horizonte									
	1	2	3	4	5	6	7	8	9	10
DYNLM	0.24	0.45	0.78	0.95	1.14	1.18	1.47	0.95	0.88	0.80
NN	33.37	10.42	6.30	20.89	15.86	17.99	17.30	14.29	5.86	11.44
CRF	9.13	6.28	5.41	8.19	3.01	4.75	2.19	6.20	2.64	3.22

De uma forma geral, os resultados obtidos são bastante piores com os pesos do que sem estes. Ainda mais perturbador é o facto de, analisando o gráfico 4.14, se observar que o melhor desempenho para o modelo DYNLM é obtido atribuindo às cheias metade do peso dos casos normais. Os pontos de cheia são aqueles que foram definidos pela EDP, tal como se explica na secção 3.3. Comparando a figura 4.15 com a 4.12a, concluiu-se que novamente a linha da previsão está mais alta, se bem que não tão é uma subida tão linear e acentuada como na regressão de quantis.

Assim conclui-se definitivamente que as cheias presentes nos anos de treino não são representativas das cheias nos anos de teste dado que atribuir mais peso a estes exemplos prejudica a performance em avaliação e dar menos peso beneficia-a. Portanto, a aplicação de métodos que tentem adaptar-se a estas não produzem bons resultados.

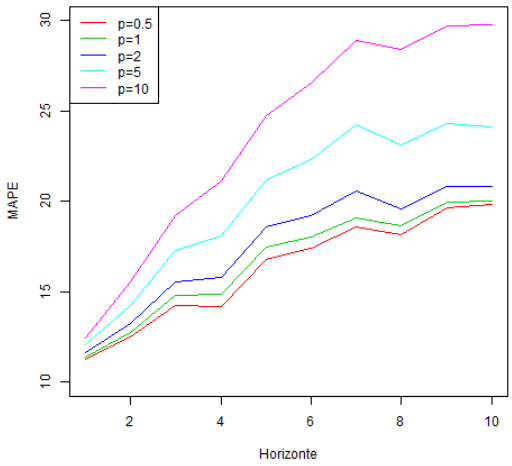


Figura 4.14: Desempenho obtido com o dynlm para vários pesos, nos casos de cheia

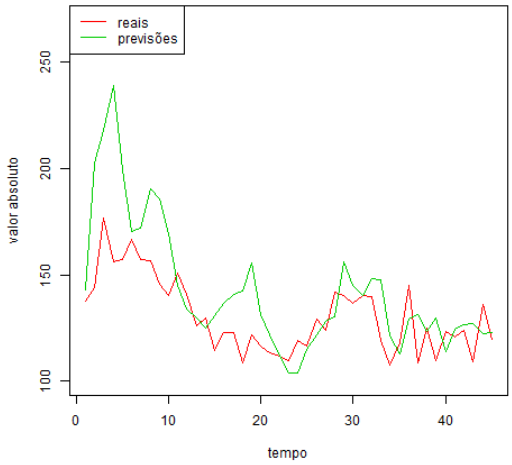


Figura 4.15: Previsões a uma hora das cheias com peso igual a 10 e valores reais

4.7 Conclusões

Ao contrário do que foi inicialmente teorizado, não foram obtidos os resultados pretendidos com nenhuma das abordagens escolhidas. De entre as várias causas que podem estar por detrás deste resultados é importante destacar:

- A granularidade dos dados pode ser demasiado grande. Dado que se pretende aproveitar o efeito que uma dada barragem mais a montante tem na que se encontra a jusante, é necessário que os dados sejam recolhidos com um período inferior ao tempo que leva a água a passar de uma para a outra.

- A quantidade de pontos de cheia pode ser demasiado pequena e as cheias de 2001 podem não ser representativas das de 2010, dado em 2001 ter ocorrido uma cheia excecional.

Com estas observações pode-se concluir que, com os dados disponíveis, a melhor hipótese de obter erros menores em avaliação será melhorar os resultados globais, na esperança que o desempenho em cheia também melhore.

Capítulo 5

Gestão de Descarregamentos: Abordagem de Simulação

Neste capítulo pretende-se apresentar uma nova abordagem ao problema da gestão de descarregamentos. Em primeiro lugar, na secção 5.1 fundamenta-se a validade e o propósito deste método, sendo os resultados obtidos expostos na secção 5.2 e 5.3. Por último, na secção 5.4, apresenta-se as conclusões obtidas através desta análise.

5.1 Descrição da abordagem

Na secção 4.7 concluiu-se que os resultados obtidos utilizando a abordagem de previsão não alcançam as metas de performance impostas pela EDP, pelos diversos motivos aí apresentados. Dado não ser actualmente possível recolher novos dados em tempo útil para este trabalho, é preciso procurar encontrar uma outra abordagem, simultaneamente válida do ponto de vista académico e útil do ponto de vista da aplicação, que utilize o conjunto de dados disponíveis.

Como referido na subsecção 2.1.2, nem todas as variáveis utilizadas têm a mesma natureza, algumas são controladas directamente pelo operador, como o caudal descarregado, outras são resultado directo das condições naturais do rio e das outras variáveis, como o caudal afluente. Do ponto de vista da aplicação, é indiferente se o valor futuro das variáveis controláveis seja previsto ou imposto, sendo que o que é realmente importante prever são as variáveis não controláveis, nomeadamente o caudal afluente. Pode-se então, na aplicação final, assumir que o valor das variáveis controláveis se mantém constante ao longo do horizonte de previsão e igual ao último valor registado, passando as variáveis controláveis a tomar valores hipotéticos em vez de previstos. No entanto, como no conjunto de dados disponíveis não se tem acesso a esse comportamento, optou-se por fazer a modelação e a avaliação considerando que o valor hipotético seguinte é igual ao valor real seguinte. Pode-se descrever a abordagem seguida no Capítulo 3 pela equação 5.1 e esta nova abordagem pela equação 5.2, no que refere à previsão das variáveis não controláveis.

$$\begin{cases} \text{se } h = 1, & \widehat{x}_{\widetilde{n}contr_t} = f\left(x_{contr_{t-p,\ldots,t-1}}, x_{\widetilde{n}contr_{t-p,\ldots,t-1}}\right) \\ \text{se } h > 1 \text{ e } h \leq p, & \widehat{x}_{\widetilde{n}contr_t} = f\left(x_{contr_{t-p,\ldots,t-h}}, \widehat{x}_{contr_{t-h+1,\ldots,t-1}}, x_{\widetilde{n}contr_{t-p,\ldots,t-h}}, \widehat{x}_{\widetilde{n}contr_{t-h+1,\ldots,t-1}}\right) \\ \text{se } h > p, & \widehat{x}_{\widetilde{n}contr_t} = f\left(\widehat{x}_{contr_{t-p,\ldots,t-1}}, \widehat{x}_{\widetilde{n}contr_{t-p,\ldots,t-1}}\right) \end{cases} \tag{5.1}$$

$$\begin{cases} \text{se } h = 1, & \widehat{x}_{\widetilde{n}contr_t} = f\left(x_{contr_{t-p,\ldots,t-1}}, x_{\widetilde{n}contr_{t-p,\ldots,t-1}}\right) \\ \text{se } h > 1 \text{ e } h \leq p, & \widehat{x}_{\widetilde{n}contr_t} = f\left(x_{contr_{t-p,\ldots,t-1}}, x_{\widetilde{n}contr_{t-p,\ldots,t-h}}, \widehat{x}_{\widetilde{n}contr_{t-h+1,\ldots,t-1}}\right) \\ \text{se } h > p, & \widehat{x}_{\widetilde{n}contr_t} = f\left(x_{contr_{t-p,\ldots,t-1}}, \widehat{x}_{\widetilde{n}contr_{t-p,\ldots,t-1}}\right) \end{cases} \tag{5.2}$$

Nestas duas equações $\widehat{x}_{contr_t}$ e $\widehat{x}_{\widetilde{n}contr_t}$ representam os valores previstos das variáveis controláveis e não controláveis, respetivamente, e x_{contr} e $x_{\widetilde{n}contr}$ os valores reais. Desta forma, em vez de construir dez modelos, um para cada variável, apenas são precisos modelos para as variáveis não controláveis.

Para $h > 1$, espera-se obter melhores resultados, dado estar-se a utilizar o valor real das variáveis controláveis em vez do previsto. No entanto, visto que o primeiro valor do caudal afluente é previsto apenas com os valores passados de todas as variáveis, o erro da previsão para $h = 1$ será igual. Sendo assim, e dado que se assume saber o valor futuro das decisões dos operadores, pode-se também acrescentar essa informação nos modelos de forma a melhorar ainda mais a sua performance. Com essa alteração a equação 5.2 passa a ser descrita da forma da equação 5.3, em que as variáveis controláveis são consideradas até ao instante t, em vez de $t-1$ como anteriormente.

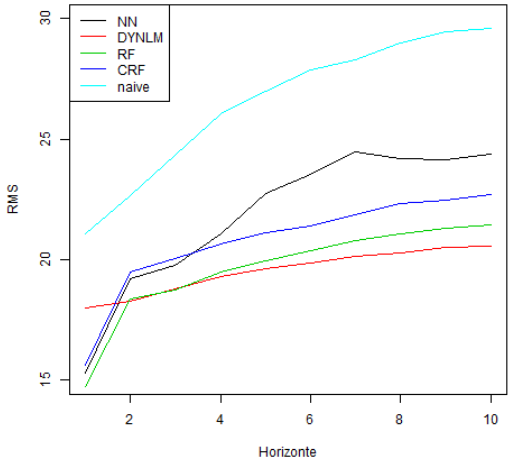
$$\begin{cases} \text{se } h = 1, & \widehat{x}_{\widetilde{n}contr_t} = f\left(x_{contr_{t-p,\ldots,t}}, x_{\widetilde{n}contr_{t-p,\ldots,t-1}}\right) \\ \text{se } h > 1 \text{ e } h \leq p, & \widehat{x}_{\widetilde{n}contr_t} = f\left(x_{contr_{t-p,\ldots,t}}, x_{\widetilde{n}contr_{t-p,\ldots,t-h}}, \widehat{x}_{\widetilde{n}contr_{t-h+1,\ldots,t-1}}\right) \\ \text{se } h > p, & \widehat{x}_{\widetilde{n}contr_t} = f\left(x_{contr_{t-p,\ldots,t}}, \widehat{x}_{\widetilde{n}contr_{t-p,\ldots,t-1}}\right) \end{cases} \tag{5.3}$$

Em seguida são apresentados os resultados obtidos com esta abordagem, no conjunto de dados inicial e no conjunto de dados com as variações, visto que este último apresentou uma melhoria de resultados na subsecção 4.3.1. Os algoritmos utilizados foram DYNLM, RF, CRF e NN, dado permitirem esta implementação.

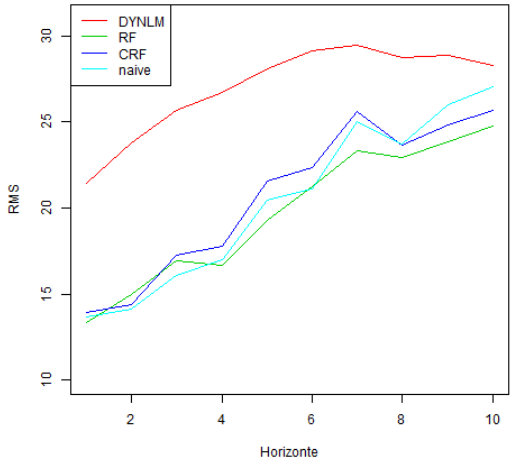
5.2 Resultados em simulação com dados originais

Considerando as previsões a uma hora, a única diferença desta implementação para as anteriores é o facto de se considerar os valores reais da próxima hora das variáveis controláveis. Por outro lado, nas restantes previsões existe a diferença significativa que, para as variáveis controláveis, os valores passados se mantêm sempre os reais, ao contrário das variáveis não controláveis que, como explicado na subsecção 2.3.2.1, tomam os valores previstos para as horas anteriores.

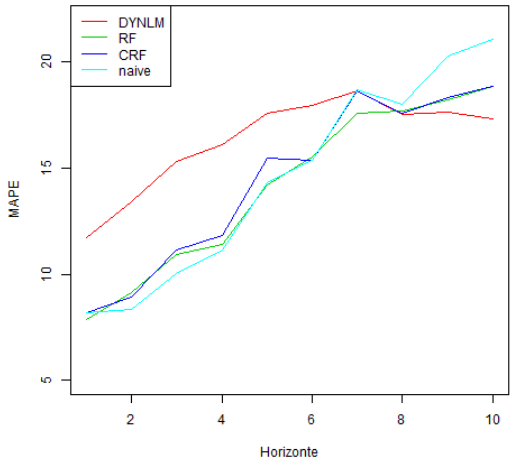
Assim, espera-se que haja uma pequena melhoria nas previsões a uma hora e que esta seja cada vez mais acentuada à medida que se vai avançando ao longo do horizonte de previsão. A figura 5.1 apresenta os resultados obtidos com esta abordagem, nomeadamente o RMSE obtido para todos os pontos e o RMSE e MAPE em cheia. Na tabela 5.1 pode-se encontrar a diferença entre o RMSE aqui apresentado e o obtido na secção 4.2 e na tabela 5.2 apresenta-se o mesmo para o MAPE.



(a) RMSE em avaliação dos modelos



(b) RMSE em avaliação dos modelos, em cheias



(c) MAPE em avaliação dos modelos, em cheias

Figura 5.1: Desempenho obtido com os algoritmos séries temporais em simulação

Da tabela 5.1 observa-se que, tal como era esperado, houve uma melhoria nos resultados para todos os modelos na avaliação do caso geral, sendo esta menor na primeira previsão e maior nas restantes. Relativamente às cheias, pode-se observar pela tabela 5.2 que também há uma melhoria nos resultados, excepto nas redes neuronais para previsões a mais de duas horas. Aliás, o erro das redes neuronais é tão elevado em cheia que nem se optou por não representar graficamente.

Tabela 5.1: Diferença entre o RMSE obtido em simulação e previsão, para todos os pontos

Modelos	Horizonte									
	1	2	3	4	5	6	7	8	9	10
DYN	0.05	-1.37	-2.41	-2.68	-2.93	-3.22	-3.33	-3.43	-3.46	-3.49
RF	-2.69	-3.42	-6.04	-5.27	-4.36	-4.05	-3.92	-3.79	-3.69	-3.68
NN	-3.46	-3.06	-7.97	-8.03	-6.93	-6.54	-7.50	-8.51	-9.65	-10.00
CRF	-2.28	-0.50	-2.94	-3.23	-3.18	-2.89	-2.84	-2.78	-3.12	-3.15

Tabela 5.2: Diferença entre o MAPE obtido em simulação e previsão, em cheia

Modelos	Horizonte									
	1	2	3	4	5	6	7	8	9	10
DYN	0.34	0.64	0.54	1.25	0.09	-0.09	-0.46	-1.08	-2.33	-2.70
RF	-0.69	-0.83	-1.11	-2.88	-3.71	-2.97	-4.34	-3.12	-3.92	-3.55
NN	-0.73	-10.65	1.53	13.57	20.79	24.02	29.10	30.03	30.00	33.58
CRF	-1.38	-0.81	-1.06	-0.46	-1.40	-1.06	-0.94	-0.04	-1.63	-1.45

Analisando o gráfico 5.1c conclui-se que, a partir das 5 horas existem vários modelos com um desempenho melhor que o modelo *naive*.

5.3 Resultados em simulação com as variações

À semelhança do que foi obtido na subsecção 4.3.1, espera-se que utilizando as variações haja uma melhoria no desempenho das previsões em cheia. De forma a averiguar se esta conclusão também se aplica na abordagem de simulação, aplicou-se o conjunto de dados das variações aos algoritmos previamente apresentados, fazendo novamente a conversão e comparação com o valor absoluto. A figura 5.2 apresenta a comparação da performance dos vários modelos utilizando as variações e as tabelas 5.3 e 5.4 a diferença entre estes resultados e os da secção 4.2.

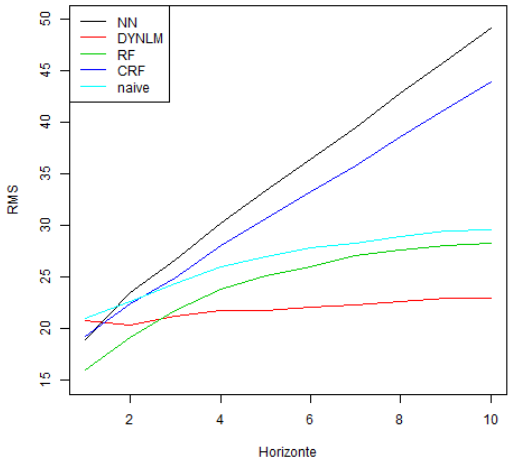
Observando as tabelas 5.3 e 5.4 pode-se concluir que, apesar de não haver uma melhoria no caso geral, o desempenho em cheia é melhor que anteriormente. Aliás, analisando o gráfico 5.2c conclui-se que, pela primeira vez, foi possível obter melhor desempenho que o modelo *naive* nas primeiras duas horas de previsão, pelo modelo RF e CRF. Nas restantes horas é de salientar o modelo DYNLM que apresenta sempre melhor performance que o modelo *naive*.

Tabela 5.3: Diferença entre o RMSE obtido em simulação e previsão, com as variações, para todos os pontos

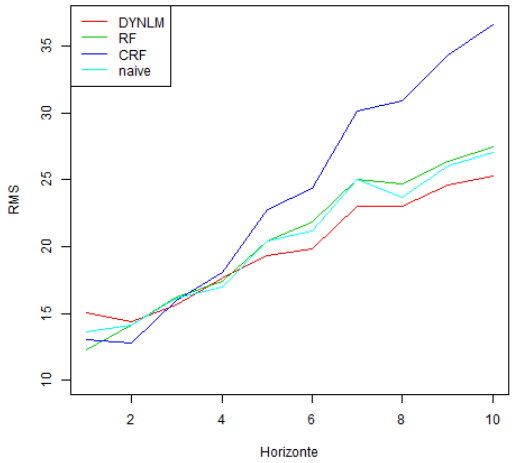
Modelos	Horizonte									
	1	2	3	4	5	6	7	8	9	10
DYN	2.88	0.72	0.04	-0.22	-0.82	-0.95	-1.13	-1.13	-0.99	-1.08
RF	-1.38	-2.60	-3.02	-0.93	0.76	1.63	2.36	2.78	3.03	3.16
NN	0.18	1.25	-1.04	1.13	3.71	6.36	7.49	10.11	12.21	14.76
CRF	1.42	2.39	1.91	4.14	6.32	8.94	11.06	13.47	15.78	18.13

Tabela 5.4: Diferença entre o MAPE obtido em simulação e previsão, com as variações, em cheia

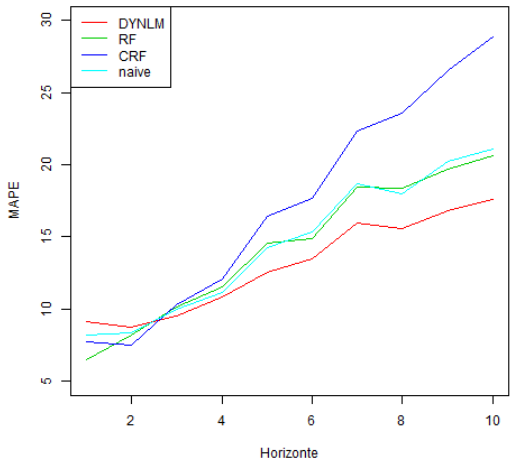
Modelos	Horizonte									
	1	2	3	4	5	6	7	8	9	10
DYN	-2.25	-4.04	-5.23	-4.05	-4.90	-4.57	-3.17	-3.06	-3.15	-2.41
RF	-2.05	-1.76	-1.87	-2.74	-3.35	-3.61	-3.51	-2.43	-2.44	-1.72
NN	-11.17	-23.90	-13.82	-5.00	-3.35	-2.44	2.61	7.95	8.54	19.09
CRF	-1.87	-2.22	-1.94	-0.16	-0.39	1.29	2.76	5.96	6.64	8.51



(a) RMSE em avaliação dos modelos



(b) RMSE em avaliação dos modelos, em cheias



(c) RMSE em avaliação dos modelos, em cheias

Figura 5.2: Desempenho obtido com os algoritmos séries temporais em simulação, com as variações

5.4 Conclusões

Com esta abordagem foi possível obter melhores resultados. Se, por exemplo, se juntar os modelos RF e DYNLM é possível obter sempre previsões melhores que as do modelo *naive*.No entanto, apesar das previsões obtidas serem úteis e apresentarem um desempenho razoável, ainda não alcançam a performance mínima estabelecida pela EDP.

Capítulo 6

Conclusões e Trabalho Futuro

Com este capítulo se encerra o estudo da gestão de descarregamentos, apresentando um breve resumo dos pontos e conclusões mais relevantes na secção 6.1, e propondo novos caminhos e soluções na secção 6.2.

6.1 Resumo e Satisfação dos Objectivos

Como foi referido na secção 1.1, os descarregamentos são indispensáveis ao bom funcionamento de uma barragem, sendo a sua gestão em épocas de cheia de elevada importância devido aos riscos inerentes. Relativamente à gestão de descarregamentos, pode-se dizer que atualmente a EDP possui um conjunto de normas [6, 7, 9] que definem quais as ações a tomar em cada situação, nomeadamente a tabela de sequência de manobras, disponível no Anexo A. Estas normas permitem ao operador determinar as aberturas das comportas, dependendo da situação atual e do regime afluente. Portanto, este é um processo bem controlado, que não apresenta grandes dificuldades exceto em situações de grandes caudais afluentes em alturas de cheias, nas quais o comportamento do rio se torna mais imprevisível. Assim, o objetivo desta tese consistiu em dar resposta a este problema, através da aplicação de técnicas de *Data Mining* a dados históricos de uma dada barragem.

Face à revisão bibliográfica apresentada no Capítulo 2, optou-se por tomar uma abordagem baseada na previsão de séries temporais, com as variantes da regressão de quantis e previsão de valores raros extremos. A primeira preocupação levantada foi o facto de haver poucos dados de cheia, sendo que apenas dois dos anos disponíveis contêm pontos desta natureza e em pouca quantidade. O impacto deste factor foi logo evidenciado na primeira iteração, na secção 4.2, onde se comprovou que os modelos apresentam bons resultados no caso geral mas não em cheia. Assim, corroborou-se a hipótese inicial de ser necessário métodos específicos para tratar as cheias, dado estas serem casos raros e extremos, validando o uso dos quantis na secção 4.5 e da previsão de valores raros extremos na secção 4.6. Para além disso, na subsecção 4.2.1 conclui-se que o tempo que a água demora da barragem de montante para a de jusante é de aproximadamente uma hora, ou seja, os dados horários têm uma granularidade muito elevada para captar este fenómeno. Ainda

assim, apesar desta dificuldade acrescida, foi possível obter resultados razoáveis no caso geral. Antes de passar para as abordagens que focam as cheias, aprofundou-se um pouco mais o estudo das séries temporais acrescentando e alterando as variáveis disponíveis. Deste estudo obteve-se melhores resultados na previsão das cheias utilizando a variação das variáveis em vez do seu valor absoluto.

Passando para a regressão de quantis, concluiu-se que este método não é adequado pois as cheias em avaliação não são previstas sempre por valores inferiores, como seria de esperar de algoritmos que se adaptam ao comportamento médio dos dados. Este estudo levou também à suspeita de que os pontos de cheia dos dados de modelação não seriam representativos dos dados de cheia de avaliação, o que faz sentido visto no primeiro caso ter ocorrido uma cheia excecional. Relativamente à previsão de valores raros extremos, mais uma vez não foram obtidos os resultados esperados. Novamente se propôs que as cheias de 2001 não são representativas das de 2010, visto que dar mais peso às cheias em modelação piora o desempenho em avaliação. Assim, conclui-se que o próximo passo seria procurar melhorar a performance global, na expectativa que as cheias também melhorassem.

Procurando aproveitar o estudo feito anteriormente e, simultaneamente, encontrar uma forma de melhorar o desempenho sem prejudicar a aplicação em causa, chegou-se à abordagem de simulação. Esta consiste em considerar que as variáveis controladas pelo operador são conhecidas à priori, não sendo por isso necessário prevê-las. Para além disso, estudou-se novamente o uso das variações *versus* o módulo das variáveis, obtendo-se os melhores resultados.

Com o finalizar deste trabalho pode-se concluir que academicamente foram alcançados todos os objetivos propostos, tendo sido ainda acrescentada uma nova abordagem na tentativa de obter melhores resultados. Pode-se assim afirmar que este estudo contribui de uma forma positiva e inovadora para o domínio da Gestão de descarregamentos, principalmente pela aplicação de novas abordagens. No entanto, devido a todos os fatores apresentados anteriormente, não foi possível alcançar o MAPE de 5% na previsão a uma hora, impostos pela EDP, sendo que o melhor que se conseguiu foi $6,48\%$ com as RF na abordagem de simulação com as variações, na secção 5.3.

6.2 Trabalho Futuro

As propostas de trabalho futuro podem ser divididas em duas grandes categorias: sugestões de trabalho e sugestões para a empresa. As sugestões de trabalho consistem noutros métodos encontrados no decorrer do projeto, que se julga que poderão melhorar os resultados. As sugestões para a empresa estão relacionadas com alguns problemas encontrados com os dados e as suas possíveis soluções.

6.2.1 Sugestões de trabalho

Relativamente ao trabalho elaborado, um próximo passo a tomar seria realizar o estudo das variáveis de forma individual. Neste trabalho, apesar de o propósito ser apenas a previsão do caudal afluente, modelou-se e previu-se todas as outras variáveis de forma a obter previsões a mais

de uma hora para a variável objetivo. Ou seja, para cada algoritmo, criaram-se dez modelos, cada um correspondente a uma das variáveis escolhidas na secção 3.5, e avaliou-se a performance obtida apenas na previsão do caudal afluente. Esta abordagem, apesar de ser mais rápida e igualmente válida dado focar a variável objetivo, peca pelo facto de poder haver algoritmos mais apropriados para uma dada variável que outros. Se fosse possível prever cada uma dessas variáveis de forma perfeita, o erro cometido na previsão do caudal afluente não iria aumentar tão acentuadamente com o horizonte de previsão. O desempenho ir-se-ia manter praticamente igual ao da previsão a uma hora, aumentando apenas pela influência do erro cometido na previsão deste, dado as outras previsões serem perfeitas. Apesar de ser impossível obter um desempenho perfeito, poder-se-ia estudar as variáveis individualmente de forma a escolher os melhores algoritmos para cada uma, obtendo um conjunto de modelos que, no global, apresentasse melhor desempenho na previsão do caudal afluente. É preciso ter atenção que, apesar de se esperar que esta abordagem melhorasse significativamente os resultados ao longo do horizonte de previsão, não influenciaria em nada as previsões a uma hora, dado que na primeira iteração os valores passados das outras variáveis são os observados.

Para além de se estudar as variáveis individualmente com os métodos propostos, poder-se-ia também averiguar novas abordagens para cada uma destas variáveis, conforme as suas características. Por exemplo, como foi referido na subsecção 4.3.2, seria interessante estudar a aplicação de métodos de *zero-inflated values*, para as variáveis que têm muitos zeros, como o caudal afluente.

Por último, alcançando o desempenho pretendido, seria necessário criar uma aplicação que fornecesse estas previsões ao operador, em tempo real e de forma apelativa. Para tal, sugere-se o desenvolvimento de uma aplicação gráfica que, com ligação à base de dados da EDP, apresentasse um ou vários perfis do caudal afluente futuro, admitindo um ou mais valores do caudal descarregado.

6.2.2 Sugestões para a empresa

As dificuldades mais graves que surgiram relacionadas com o conjunto de dados fornecido foram a falta de pontos de cheia e a granularidade destes ser demasiado grossa. Quanto ao primeiro problema, a única forma de o solucionar seria procurar cheias mais antigas e, dado que antes de 2001 os dados não eram recolhidos de forma automática, fazer um tratamento e limpeza cuidados a esses. Desta forma, aumentar-se-ia o número de dados de cheia, tanto em modelação como em avaliação, permitindo um estudo e conclusões mais sólidas. Para além disso, possivelmente seria mais útil treinar os modelos com dados de cheias mais típicas em vez dos dados da cheia excecional de 2001, dado a frequência desta última ser muito mais baixa.

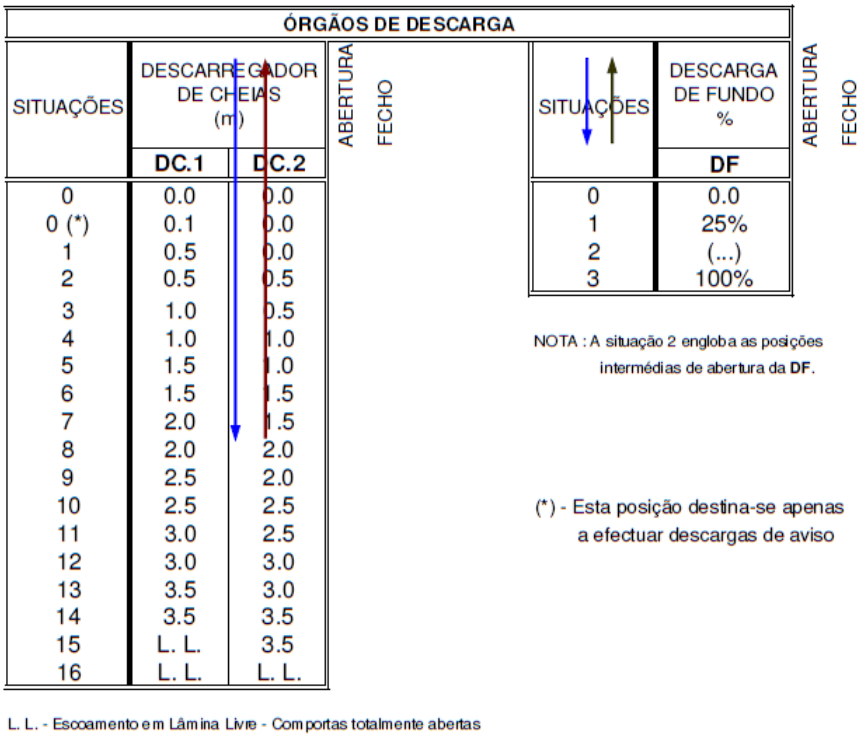
Relativamente ao problema da granularidade, este apenas poderia ser solucionado com uma nova recolha de dados, passando estes a ser recolhidos a uma frequência apropriada para a distância entre a barragem em questão e as que se encontram imediatamente a montante e jusante desta. Alternativamente, poderia-se optar por manter uma abordagem como na secção 4.4, derivando o conjunto de dados necessário dos dados esporádicos, dado que estes contêm os registos das variáveis sempre que existe uma alteração significativa no seu valor. Para este caso, seria necessário

que todas as variáveis fossem incluídas no conjunto de dados esporádicos, principalmente o caudal afluente visto ser a variável objetivo. Prevê-se que, mesmo com esta segunda abordagem, a alteração da granularidade dos dados seja um factor predominante na obtenção de bons resultados, no que diz respeito à previsão das variáveis hidrológicas.

Anexo A

Auxiliares normativos

A.1 Sequência de Manobras de Venda Nova



ÓRGÃOS DE DESCARGA

SITUAÇÕES	DESCARREGADOR DE CHEIAS (m)		ABERTURA / FECHO
	DC.1	DC.2	
0	0.0	0.0	
0 (*)	0.1	0.0	
1	0.5	0.0	
2	0.5	0.5	
3	1.0	0.5	
4	1.0	1.0	
5	1.5	1.0	
6	1.5	1.5	
7	2.0	1.5	
8	2.0	2.0	
9	2.5	2.0	
10	2.5	2.5	
11	3.0	2.5	
12	3.0	3.0	
13	3.5	3.0	
14	3.5	3.5	
15	L. L.	3.5	
16	L. L.	L. L.	

SITUAÇÕES	DESCARGA DE FUNDO %	ABERTURA / FECHO
	DF	
0	0.0	
1	25%	
2	(...)	
3	100%	

NOTA : A situação 2 engloba as posições intermédias de abertura da **DF**.

(*) - Esta posição destina-se apenas a efectuar descargas de aviso

L. L. - Escoamento em Lâmina Livre - Comportas totalmente abertas

Figura A.1: Sequência de Manobras de Salamonde [9]

A.2 Curvas de Regolfo a Montante da Barragem de Pocinho

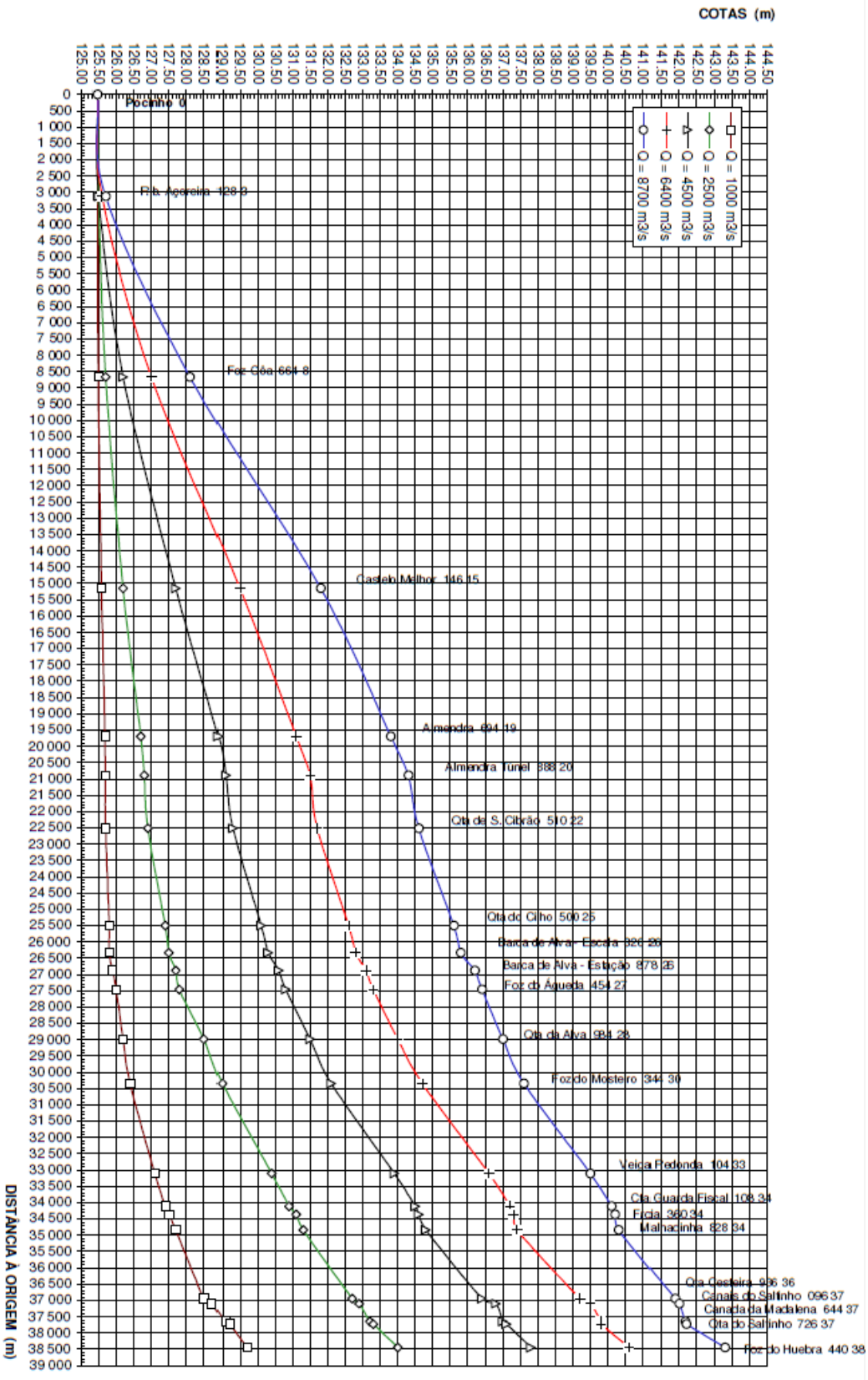


Figura A.2: Curvas de Regolfo a Montante da Barragem de Pocinho [6]

A.3 Tabela dos volumes armazenados na albufeira

	COTAS (m)	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
NME	700.50	96.37									
	700.40	96.00	96.03	96.07	96.11	96.15	96.18	96.22	96.26	96.30	96.33
	700.30	95.62	95.66	95.70	95.73	95.77	95.81	95.85	95.88	95.92	95.96
	700.20	95.25	95.29	95.32	95.36	95.40	95.44	95.47	95.51	95.55	95.59
	700.10	94.87	94.91	94.95	94.99	95.02	95.06	95.10	95.14	95.17	95.21
NPA	700.00	94.50	94.54	94.57	94.61	94.65	94.69	94.72	94.76	94.80	94.84
	699.90	94.13	94.16	94.20	94.24	94.28	94.31	94.35	94.39	94.43	94.46
	699.80	93.75	93.79	93.83	93.86	93.90	93.94	93.98	94.01	94.05	94.09
	699.70	93.38	93.41	93.45	93.49	93.53	93.56	93.60	93.64	93.68	93.71
	699.60	93.00	93.04	93.08	93.11	93.15	93.19	93.23	93.26	93.30	93.34
	699.50	92.63	92.66	92.70	92.74	92.78	92.81	92.85	92.89	92.93	92.96
	699.40	92.25	92.29	92.33	92.36	92.40	92.44	92.48	92.51	92.55	92.59
	699.30	91.88	91.92	91.95	91.99	92.03	92.07	92.10	92.14	92.18	92.22
	699.20	91.50	91.54	91.58	91.62	91.65	91.69	91.73	91.77	91.80	91.84
	699.10	91.13	91.17	91.20	91.24	91.28	91.32	91.35	91.39	91.43	91.47
	699.00	90.76	90.79	90.83	90.87	90.91	90.94	90.98	91.02	91.05	91.09
	698.90	90.38	90.42	90.46	90.49	90.53	90.57	90.61	90.64	90.68	90.72
	698.80	90.01	90.05	90.08	90.12	90.16	90.20	90.23	90.27	90.31	90.34
	698.70	89.64	89.67	89.71	89.75	89.78	89.82	89.86	89.90	89.93	89.97
	698.60	89.26	89.30	89.34	89.37	89.41	89.45	89.49	89.52	89.56	89.60
	698.50	88.89	88.93	88.97	89.00	89.04	89.08	89.11	89.15	89.19	89.23
	698.40	88.52	88.56	88.59	88.63	88.67	88.70	88.74	88.78	88.82	88.85
	698.30	88.15	88.18	88.22	88.26	88.30	88.33	88.37	88.41	88.44	88.48
	698.20	87.78	87.81	87.85	87.89	87.92	87.96	88.00	88.04	88.07	88.11
	698.10	87.41	87.44	87.48	87.52	87.55	87.59	87.63	87.66	87.70	87.74
	698.00	87.04	87.07	87.11	87.15	87.18	87.22	87.26	87.29	87.33	87.37
	697.90	86.67	86.70	86.74	86.78	86.81	86.85	86.89	86.92	86.96	87.00
	697.80	86.30	86.33	86.37	86.41	86.44	86.48	86.52	86.56	86.59	86.63
	697.70	85.93	85.97	86.00	86.04	86.08	86.11	86.15	86.19	86.22	86.26
	697.60	85.56	85.60	85.63	85.67	85.71	85.75	85.78	85.82	85.86	85.89
	697.50	85.19	85.23	85.27	85.30	85.34	85.38	85.41	85.45	85.49	85.52
	697.40	84.83	84.86	84.90	84.94	84.97	85.01	85.05	85.08	85.12	85.16
	697.30	84.46	84.50	84.54	84.57	84.61	84.65	84.68	84.72	84.75	84.79
	697.20	84.10	84.13	84.17	84.21	84.24	84.28	84.32	84.35	84.39	84.43
	697.10	83.73	83.77	83.81	83.84	83.88	83.92	83.95	83.99	84.02	84.06
	697.00	83.37	83.41	83.44	83.48	83.52	83.55	83.59	83.62	83.66	83.70
	696.90	83.01	83.04	83.08	83.12	83.15	83.19	83.23	83.26	83.30	83.33
	696.80	82.65	82.68	82.72	82.75	82.79	82.83	82.86	82.90	82.94	82.97
	696.70	82.29	82.32	82.36	82.39	82.43	82.47	82.50	82.54	82.57	82.61
	696.60	81.93	81.96	82.00	82.03	82.07	82.11	82.14	82.18	82.21	82.25
	696.50	81.57	81.60	81.64	81.67	81.71	81.75	81.78	81.82	81.85	81.89
	696.40	81.21	81.24	81.28	81.32	81.35	81.39	81.42	81.46	81.49	81.53
	696.30	80.85	80.89	80.92	80.96	80.99	81.03	81.07	81.10	81.14	81.17
	696.20	80.49	80.53	80.57	80.60	80.64	80.67	80.71	80.74	80.78	80.82
	696.10	80.14	80.17	80.21	80.25	80.28	80.32	80.35	80.39	80.42	80.46
	696.00	79.79	79.82	79.86	79.89	79.93	79.96	80.00	80.03	80.07	80.10
	695.90	79.43	79.47	79.50	79.54	79.57	79.61	79.64	79.68	79.71	79.75
	695.80	79.08	79.11	79.15	79.19	79.22	79.26	79.29	79.33	79.36	79.40
	695.70	78.73	78.76	78.80	78.83	78.87	78.90	78.94	78.97	79.01	79.04
	695.60	78.38	78.41	78.45	78.48	78.52	78.55	78.59	78.62	78.66	78.69
	695.50	78.03	78.06	78.10	78.13	78.17	78.20	78.24	78.27	78.31	78.34
	695.40	77.68	77.72	77.75	77.79	77.82	77.85	77.89	77.92	77.96	77.99
	695.30	77.33	77.37	77.40	77.44	77.47	77.51	77.54	77.58	77.61	77.65
	695.20	76.99	77.02	77.06	77.09	77.13	77.16	77.20	77.23	77.26	77.30
	695.10	76.64	76.68	76.71	76.75	76.78	76.82	76.85	76.88	76.92	76.95
	695.00	76.30	76.33	76.37	76.40	76.44	76.47	76.51	76.54	76.57	76.61
	694.90	75.96	75.99	76.03	76.06	76.09	76.13	76.16	76.20	76.23	76.27
	694.80	75.62	75.65	75.68	75.72	75.75	75.79	75.82	75.86	75.89	75.92
	694.70	75.28	75.31	75.34	75.38	75.41	75.45	75.48	75.51	75.55	75.58
	694.60	74.94	74.97	75.01	75.04	75.07	75.11	75.14	75.17	75.21	75.24
	694.50	74.60	74.63	74.67	74.70	74.73	74.77	74.80	74.84	74.87	74.90

Figura A.3: Tabela dos valores mais altos de volume armazenado na albufeira de Venda Nova [8]

COTAS (m)	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90
629.00	0.26	0.27	0.27	0.28	0.28	0.29	0.29	0.30	0.30	0.31
628.00	0.22	0.23	0.23	0.24	0.24	0.24	0.25	0.25	0.26	0.26
627.00	0.19	0.19	0.20	0.20	0.20	0.21	0.21	0.21	0.22	0.22
626.00	0.16	0.16	0.16	0.17	0.17	0.17	0.18	0.18	0.18	0.19
625.00	0.13	0.13	0.14	0.14	0.14	0.14	0.15	0.15	0.15	0.15
624.00	0.11	0.11	0.11	0.11	0.11	0.12	0.12	0.12	0.12	0.13
623.00	0.08	0.09	0.09	0.09	0.09	0.09	0.10	0.10	0.10	0.10
622.00	0.06	0.07	0.07	0.07	0.07	0.07	0.07	0.08	0.08	0.08
621.00	0.05	0.05	0.05	0.05	0.05	0.05	0.06	0.06	0.06	0.06
620.00	0.03	0.03	0.03	0.03	0.04	0.04	0.04	0.04	0.04	0.04
619.00	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.03	0.03	0.03
618.00	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02
617.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01

Df → 610.70
Nf → 610.00

Figura A.4: Tabela dos valores mais baixos de volume armazenado na albufeira de Venda Nova [8]

	COTAS (m)	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
NME	880.10	571.08									
NPA	880.00	568.69	568.93	569.17	569.41	569.65	569.88	570.12	570.36	570.60	570.84
	879.90	566.31	566.55	566.78	567.02	567.26	567.50	567.74	567.97	568.21	568.45
	879.80	563.94	564.18	564.41	564.65	564.89	565.12	565.36	565.60	565.83	566.07
	879.70	561.58	561.82	562.05	562.29	562.52	562.76	563.00	563.23	563.47	563.70
	879.60	559.23	559.47	559.70	559.94	560.17	560.41	560.64	560.88	561.11	561.35
	879.50	556.90	557.13	557.36	557.60	557.83	558.06	558.30	558.53	558.77	559.00
	879.40	554.57	554.80	555.04	555.27	555.50	555.73	555.97	556.20	556.43	556.66
	879.30	552.26	552.49	552.72	552.95	553.18	553.41	553.64	553.88	554.11	554.34
	879.20	549.95	550.18	550.41	550.64	550.87	551.10	551.33	551.56	551.79	552.03
	879.10	547.66	547.89	548.12	548.34	548.57	548.80	549.03	549.26	549.49	549.72
	879.00	545.37	545.60	545.83	546.06	546.29	546.51	546.74	546.97	547.20	547.43
	878.90	543.10	543.33	543.55	543.78	544.01	544.24	544.46	544.69	544.92	545.15
	878.80	540.84	541.06	541.29	541.52	541.74	541.97	542.19	542.42	542.65	542.87
	878.70	538.58	538.81	539.03	539.26	539.48	539.71	539.93	540.16	540.39	540.61
	878.60	536.34	536.56	536.79	537.01	537.24	537.46	537.69	537.91	538.13	538.36
	878.50	534.11	534.33	534.55	534.78	535.00	535.22	535.45	535.67	535.89	536.12
	878.40	531.88	532.11	532.33	532.55	532.77	532.99	533.22	533.44	533.66	533.88
	878.30	529.67	529.89	530.11	530.33	530.55	530.78	531.00	531.22	531.44	531.66
	878.20	527.47	527.69	527.91	528.13	528.35	528.57	528.79	529.01	529.23	529.45
	878.10	525.27	525.49	525.71	525.93	526.15	526.37	526.59	526.81	527.03	527.25
	878.00	523.09	523.30	523.52	523.74	523.96	524.18	524.40	524.62	524.83	525.05
	877.90	520.91	521.13	521.35	521.56	521.78	522.00	522.22	522.43	522.65	522.87
	877.80	518.74	518.96	519.18	519.39	519.61	519.83	520.04	520.26	520.48	520.69
	877.70	516.59	516.80	517.02	517.23	517.45	517.66	517.88	518.10	518.31	518.53
	877.60	514.44	514.65	514.87	515.08	515.30	515.51	515.73	515.94	516.16	516.37
	877.50	512.30	512.51	512.73	512.94	513.15	513.37	513.58	513.80	514.01	514.22
	877.40	510.17	510.38	510.59	510.81	511.02	511.23	511.45	511.66	511.87	512.09
	877.30	508.05	508.26	508.47	508.68	508.90	509.11	509.32	509.53	509.74	509.96
	877.20	505.93	506.15	506.36	506.57	506.78	506.99	507.20	507.41	507.62	507.84
	877.10	503.83	504.04	504.25	504.46	504.67	504.88	505.09	505.30	505.51	505.72
	877.00	501.74	501.94	502.15	502.36	502.57	502.78	502.99	503.20	503.41	503.62
	876.90	499.65	499.86	500.07	500.27	500.48	500.69	500.90	501.11	501.32	501.53
	876.80	497.57	497.78	497.98	498.19	498.40	498.61	498.82	499.02	499.23	499.44
	876.70	495.50	495.71	495.91	496.12	496.33	496.53	496.74	496.95	497.16	497.36
	876.60	493.44	493.64	493.85	494.06	494.26	494.47	494.67	494.88	495.09	495.29
	876.50	491.38	491.59	491.79	492.00	492.20	492.41	492.61	492.82	493.03	493.23
	876.40	489.34	489.54	489.75	489.95	490.16	490.36	490.56	490.77	490.97	491.18
	876.30	487.30	487.50	487.71	487.91	488.11	488.32	488.52	488.73	488.93	489.13
	876.20	485.27	485.47	485.68	485.88	486.08	486.28	486.49	486.69	486.89	487.10
	876.10	483.25	483.45	483.65	483.85	484.06	484.26	484.46	484.66	484.86	485.07
	876.00	481.23	481.43	481.64	481.84	482.04	482.24	482.44	482.64	482.84	483.05
	875.90	479.23	479.43	479.63	479.83	480.03	480.23	480.43	480.63	480.83	481.03
	875.80	477.23	477.43	477.63	477.83	478.03	478.23	478.43	478.63	478.83	479.03
	875.70	475.24	475.44	475.63	475.83	476.03	476.23	476.43	476.63	476.83	477.03
	875.60	473.25	473.45	473.65	473.85	474.05	474.24	474.44	474.64	474.84	475.04
	875.50	471.28	471.47	471.67	471.87	472.07	472.26	472.46	472.66	472.86	473.05
	875.40	469.31	469.50	469.70	469.90	470.09	470.29	470.49	470.68	470.88	471.08
	875.30	467.34	467.54	467.74	467.93	468.13	468.32	468.52	468.72	468.91	469.11
	875.20	465.39	465.58	465.78	465.97	466.17	466.37	466.56	466.76	466.95	467.15
	875.10	463.44	463.64	463.83	464.02	464.22	464.41	464.61	464.80	465.00	465.19
	875.00	461.50	461.69	461.89	462.08	462.28	462.47	462.66	462.86	463.05	463.25
	874.90	459.57	459.76	459.95	460.15	460.34	460.53	460.73	460.92	461.11	461.31
	874.80	457.64	457.83	458.02	458.22	458.41	458.60	458.79	458.99	459.18	459.37
	874.70	455.72	455.91	456.10	456.29	456.49	456.68	456.87	457.06	457.25	457.45
	874.60	453.81	454.00	454.19	454.38	454.57	454.76	454.95	455.14	455.34	455.53
	874.50	451.90	452.09	452.28	452.47	452.66	452.85	453.04	453.23	453.42	453.61
	874.40	450.00	450.19	450.38	450.57	450.76	450.95	451.14	451.33	451.52	451.71
	874.30	448.10	448.29	448.48	448.67	448.86	449.05	449.24	449.43	449.62	449.81
	874.20	446.22	446.41	446.59	446.78	446.97	447.16	447.35	447.54	447.73	447.92
	874.10	444.34	444.52	444.71	444.90	445.09	445.28	445.46	445.65	445.84	446.03
	874.00	442.46	442.65	442.84	443.02	443.21	443.40	443.59	443.77	443.96	444.15

Figura A.5: Tabela dos valores mais altos de volume armazenado na albufeira de Alto Rabagão [5]

	COTAS (m)	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90
	826.00	6.92	6.93	6.94	6.95	6.96	6.97	6.98	7.00	7.01	7.02
	825.00	5.94	5.95	5.96	5.97	5.98	5.99	6.00	6.00	6.01	6.02
	824.00	5.09	5.09	5.10	5.11	5.12	5.13	5.13	5.14	5.15	5.16
	823.00	4.34	4.35	4.36	4.36	4.37	4.38	4.39	4.39	4.40	4.41
	822.00	3.70	3.71	3.72	3.72	3.73	3.73	3.74	3.75	3.75	3.76
	821.00	3.15	3.16	3.16	3.17	3.17	3.18	3.18	3.19	3.19	3.20
	820.00	2.68	2.68	2.69	2.69	2.70	2.70	2.71	2.71	2.72	2.72
	819.00	2.28	2.28	2.28	2.29	2.29	2.29	2.30	2.30	2.31	2.31
	818.00	1.93	1.93	1.94	1.94	1.94	1.95	1.95	1.95	1.96	1.96
	817.00	1.64	1.64	1.64	1.64	1.65	1.65	1.65	1.66	1.66	1.66
	816.00	1.39	1.39	1.39	1.39	1.39	1.40	1.40	1.40	1.40	1.41
	815.00	1.17	1.17	1.17	1.18	1.18	1.18	1.18	1.18	1.19	1.19
	814.00	0.99	0.99	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00
	813.00	0.83	0.83	0.83	0.83	0.83	0.83	0.84	0.84	0.84	0.84
	812.00	0.69	0.69	0.69	0.69	0.69	0.69	0.70	0.70	0.70	0.70
	811.00	0.57	0.57	0.57	0.57	0.57	0.57	0.57	0.58	0.58	0.58
	810.00	0.46	0.46	0.46	0.46	0.46	0.47	0.47	0.47	0.47	0.47
	809.00	0.37	0.37	0.37	0.37	0.37	0.37	0.37	0.37	0.37	0.37
	808.00	0.28	0.28	0.28	0.28	0.29	0.29	0.29	0.29	0.29	0.29
	807.00	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.22	0.22
	806.00	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15
	805.00	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10
	804.00	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09
	803.00	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07
	802.00	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06
	801.00	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04
	800.00	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03
	799.00	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
Df	798.00	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	797.00	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	796.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	795.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	794.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Nf	793.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Figura A.6: Tabela dos valores mais baixos de volume armazenado na albufeira de Alto Rabagão [5]

Anexo B

Resultados

Tabela B.1: RMSE obtido na fase de modelação para os algoritmos de séries temporais

Modelos	Horizonte									
	1	2	3	4	5	6	7	8	9	10
VAR	13.71	16.68	19.38	21.87	23.76	25.43	26.74	27.61	28.12	28.48
DYNLM	13.44	16.52	18.72	20.29	21.17	22.00	22.67	23.18	23.67	24.06
BVAR	14.83	16.97	19.08	20.63	21.39	22.22	22.90	23.37	23.77	24.03
RF	14.13	22.94	29.49	29.56	29.78	30.35	30.64	31.39	31.68	32.13
NN	13.04	16.33	24.72	25.54	23.89	24.67	24.95	25.67	26.07	26.22
CRF	13.30	16.30	19.36	22.97	23.83	23.61	23.81	23.99	24.13	24.29
naive	16.07	18.48	20.98	23.56	24.68	26.11	27.41	28.27	29.18	29.72

Tabela B.2: RMSE obtido na fase de avaliação para os algoritmos de séries temporais

Modelos	Horizonte									
	1	2	3	4	5	6	7	8	9	10
VAR	18.30	20.21	23.23	25.42	27.81	29.64	31.23	32.38	33.21	33.79
DYNLM	17.94	19.65	21.19	21.97	22.55	23.08	23.45	23.73	23.95	24.03
BVAR	18.58	20.18	21.49	22.66	23.42	23.97	24.50	24.87	25.14	25.25
RF	17.40	21.77	24.77	24.74	24.33	24.41	24.72	24.86	25.00	25.13
NN	18.74	22.25	27.74	29.12	29.65	30.08	31.98	32.71	33.81	34.38
CRF	17.88	20.00	22.99	23.89	24.31	24.30	24.72	25.12	25.57	25.83
naive	21.05	22.67	24.35	26.05	26.99	27.85	28.28	28.99	29.45	29.62

Tabela B.3: RMSE obtido na fase de avaliação para os algoritmos de séries temporais, em cheias

Modelos	Horizonte									
	1	2	3	4	5	6	7	8	9	10
VAR	21.20	22.36	26.02	27.27	33.00	38.29	42.35	45.07	48.34	50.44
DYNLM	22.06	24.88	26.57	25.71	27.87	29.51	30.78	30.16	31.02	30.68
BVAR	12.32	17.90	19.86	22.59	24.44	26.97	30.60	31.84	31.60	33.66
RF	14.28	16.21	18.71	21.61	24.64	25.90	28.76	27.45	29.32	30.31
NN	49.54	63.04	67.89	55.34	58.84	56.22	69.53	72.17	68.50	60.37
CRF	15.35	15.74	19.51	18.66	23.10	23.32	26.62	23.95	26.77	27.65
naive	13.64	14.13	16.10	17.00	20.43	21.14	25.03	23.73	26.02	27.03

Tabela B.4: MAPE obtido na fase de avaliação para os algoritmos de séries temporais, em cheias

Modelos	Horizonte									
	1	2	3	4	5	6	7	8	9	10
VAR	11.75	12.69	14.93	16.83	21.25	25.01	28.65	31.65	34.10	36.80
DYNLM	11.37	12.77	14.77	14.87	17.47	18.05	19.11	18.63	19.95	20.03
BVAR	7.17	10.69	12.54	14.40	15.54	17.44	20.33	21.14	21.03	22.72
RF	8.53	9.97	12.02	14.31	17.90	18.50	21.93	20.82	22.14	22.39
NN	20.47	35.85	30.70	26.20	29.33	32.22	33.77	32.36	36.50	29.65
CRF	9.59	9.70	12.22	12.27	16.85	16.42	19.60	17.62	19.93	20.33
naive	8.19	8.32	10.01	11.12	14.28	15.36	18.69	18.01	20.27	21.06

Tabela B.5: RMSE obtido na fase de avaliação para os algoritmos de séries temporais sem recursividade

Modelos	Horizonte									
	1	2	3	4	5	6	7	8	9	10
DYNLM	17.94	21.79	23.25	24.12	25.36	24.60	24.41	24.60	25.11	24.88
RF	21.68	24.46	26.40	28.00	29.37	28.92	29.46	31.38	30.36	31.59
NN	20.27	22.48	24.12	24.93	25.85	25.98	26.12	26.44	26.63	27.00
CRF	20.41	23.08	25.07	25.87	26.74	27.10	27.47	27.84	27.97	28.31
naive	21.05	22.67	24.35	26.05	26.99	27.85	28.28	28.99	29.45	29.62

Tabela B.6: RMSE obtido na fase de avaliação para os algoritmos de séries temporais sem recursividade, em cheias

Modelos	Horizonte									
	1	2	3	4	5	6	7	8	9	10
DYNLM	15.63	13.70	18.36	22.98	26.76	27.46	29.91	30.61	35.20	42.04
RF	21.19	26.34	31.10	46.85	41.52	38.56	38.69	49.66	43.60	51.32
NN	21.45	22.36	23.87	27.80	28.27	34.86	38.14	40.82	43.62	45.52
CRF	16.61	19.04	21.81	23.47	24.66	27.24	32.36	37.66	40.02	43.30
naive	13.64	14.13	16.10	17.00	20.43	21.14	25.03	23.73	26.02	27.03

Tabela B.7: MAPE obtido na fase de avaliação para os algoritmos de séries temporais sem recursividade, em cheias

Modelos	Horizonte									
	1	2	3	4	5	6	7	8	9	10
DYNLM	11.08	9.93	12.90	16.68	20.54	20.77	25.59	25.94	26.46	33.44
RF	12.04	15.27	16.96	24.57	24.33	21.69	23.32	29.81	27.54	34.80
NN	13.76	14.28	16.03	18.61	19.79	23.44	24.44	25.35	27.13	29.23
CRF	10.44	12.61	13.64	14.33	14.49	16.55	18.52	21.85	22.50	25.88
naive	8.19	8.32	10.01	11.12	14.28	15.36	18.69	18.01	20.27	21.06

Tabela B.8: RMSE obtido na fase de avaliação para os algoritmos de séries temporais com as variações

Modelos	Horizonte									
	1	2	3	4	5	6	7	8	9	10
VAR	18.19	20.10	21.89	22.95	23.70	24.43	24.87	25.25	25.51	25.54
DYNLM	18.24	20.13	21.97	22.92	23.68	24.34	24.76	25.14	25.38	25.43
BVAR	18.68	20.66	22.76	24.00	25.07	25.86	26.44	27.07	27.45	27.55
RF	19.20	22.58	26.11	26.91	28.30	30.37	32.45	34.15	36.27	37.96
NN	18.64	21.09	23.52	24.68	25.81	26.73	27.54	28.01	28.26	28.34
CRF	19.37	21.40	24.32	24.98	26.37	28.06	28.96	30.34	30.97	31.97
naive	21.05	22.67	24.35	26.05	26.99	27.85	28.28	28.99	29.45	29.62

Tabela B.9: RMSE obtido na fase de avaliação para os algoritmos de séries temporais com as variações, em cheias

Modelos	Horizonte									
	1	2	3	4	5	6	7	8	9	10
VAR	12.56	15.21	17.64	17.73	20.74	22.61	26.31	25.71	27.43	28.04
DYNLM	16.02	17.82	22.20	21.10	24.71	26.40	29.67	29.51	31.37	31.85
BVAR	13.89	13.10	14.87	15.31	18.73	20.49	23.47	23.19	25.33	25.94
RF	13.93	21.33	21.71	20.72	22.68	23.99	25.73	26.54	29.67	33.42
NN	15.75	17.78	17.74	18.50	22.72	23.63	26.48	26.92	27.66	28.78
CRF	11.42	15.99	18.19	17.52	18.72	17.56	21.19	20.07	19.37	19.26
naive	13.64	14.13	16.10	17.00	20.43	21.14	25.03	23.73	26.02	27.03

Tabela B.10: MAPE obtido na fase de avaliação para os algoritmos de séries temporais com as variações, em cheias

Modelos	Horizonte									
	1	2	3	4	5	6	7	8	9	10
VAR	7.55	8.64	10.14	10.92	13.34	14.72	17.71	17.11	18.02	19.11
DYNLM	9.41	10.00	11.89	12.11	14.83	16.37	19.04	18.57	19.69	20.74
BVAR	7.98	7.60	9.35	10.04	13.37	14.90	17.70	17.37	19.65	20.75
RF	7.25	14.11	13.96	12.83	15.16	15.67	17.39	19.00	20.29	24.12
NN	8.97	10.92	11.10	12.26	15.36	15.90	19.09	19.09	20.00	21.60
CRF	6.80	9.05	11.14	11.08	11.96	11.63	14.84	13.93	13.56	13.87
naive	8.19	8.32	10.01	11.12	14.28	15.36	18.69	18.01	20.27	21.06

Tabela B.11: RMSE obtido na fase de avaliação para os algoritmos de séries temporais com o último valor registado

Modelos	Horizonte									
	1	2	3	4	5	6	7	8	9	10
VAR	16.98	19.73	21.35	22.64	23.76	24.55	25.51	26.25	26.94	27.51
DYNLM	16.53	19.40	21.01	22.11	22.86	23.35	23.87	24.20	24.44	24.57
BVAR	18.57	20.31	21.69	22.95	23.78	24.42	25.01	25.44	25.75	25.92
RF	16.36	19.82	22.43	23.25	24.17	24.76	25.44	25.99	26.40	26.65
NN	19.01	22.27	43.77	26.99	27.10	34.84	26.60	26.71	30.41	28.54
CRF	17.22	20.29	22.09	23.15	24.07	24.92	25.74	26.51	26.96	27.46
naive	21.05	22.67	24.35	26.05	26.99	27.85	28.28	28.99	29.45	29.62

Tabela B.12: RMSE obtido na fase de avaliação para os algoritmos de séries temporais com o último valor registado, em cheias

Modelos	Horizonte									
	1	2	3	4	5	6	7	8	9	10
VAR	21.43	25.23	27.06	25.86	26.13	28.83	29.17	29.45	30.15	29.74
DYNLM	23.15	27.67	31.08	31.36	32.08	33.55	33.29	32.23	31.75	30.87
BVAR	12.13	17.58	19.26	21.60	23.05	25.14	28.51	29.06	28.14	29.53
RF	14.07	17.59	20.55	24.06	27.21	29.50	32.00	31.13	31.56	32.02
NN	88.61	109.82	124.31	131.55	106.14	84.39	65.15	77.12	63.62	57.54
CRF	14.16	15.76	17.87	17.96	20.87	22.77	23.48	21.32	22.79	22.43
naive	13.64	14.13	16.10	17.00	20.43	21.14	25.03	23.73	26.02	27.03

Tabela B.13: MAPE obtido na fase de avaliação para os algoritmos de séries temporais com o último valor registado, em cheias

Modelos	Horizonte									
	1	2	3	4	5	6	7	8	9	10
VAR	11.75	13.63	14.58	14.55	16.18	17.47	18.26	17.74	19.63	19.82
DYNLM	12.00	14.58	16.62	17.28	18.87	19.54	20.03	19.25	19.29	18.92
BVAR	6.86	10.44	11.93	13.70	14.64	16.26	18.98	18.88	17.84	20.00
RF	8.93	11.61	13.84	17.03	20.56	21.91	24.74	23.96	24.13	24.75
NN	43.90	49.82	58.12	72.21	61.16	48.23	43.51	47.54	37.16	40.65
CRF	8.59	8.84	10.26	11.09	14.15	14.73	16.22	14.79	15.59	15.30
naive	8.19	8.32	10.01	11.12	14.28	15.36	18.69	18.01	20.27	21.06

Tabela B.14: RMSE obtido na fase de avaliação para os algoritmos de séries temporais com a variância

Modelos	Horizonte									
	1	2	3	4	5	6	7	8	9	10
VAR	18.33	20.38	23.54	25.87	28.25	29.97	31.40	32.39	33.10	33.58
DYNLM	17.96	19.70	21.23	22.01	22.57	23.10	23.46	23.73	23.94	24.01
BVAR	49.15	62.04	68.69	61.18	60.18	55.99	54.19	52.19	50.79	49.43
RF	16.72	19.73	22.62	22.75	23.04	23.33	23.79	24.24	24.48	24.72
NN	21.00	24.36	29.72	31.75	31.37	31.10	31.79	31.41	31.24	32.29
CRF	17.44	19.71	21.46	22.39	22.94	23.40	23.95	24.41	24.56	24.71
naive	21.05	22.67	24.35	26.05	26.99	27.85	28.28	28.99	29.45	29.62

Tabela B.15: RMSE obtido na fase de avaliação para os algoritmos de séries temporais com a variância, em cheias

Modelos	Horizonte									
	1	2	3	4	5	6	7	8	9	10
VAR	21.35	22.41	26.67	28.20	33.35	38.53	41.94	44.25	46.80	48.19
DYNLM	22.77	25.08	27.16	26.45	28.12	29.82	30.76	30.14	30.76	30.29
BVAR	26.82	19.66	25.07	29.04	33.10	37.99	43.16	47.33	50.02	53.49
RF	15.22	18.25	22.33	27.00	30.92	34.18	36.85	37.96	39.70	41.60
NN	92.81	67.68	82.56	85.95	83.22	49.79	63.84	77.51	86.46	103.78
CRF	15.46	15.10	17.79	17.50	21.33	22.14	24.90	22.56	23.73	23.03
naive	13.64	14.13	16.10	17.00	20.43	21.14	25.03	23.73	26.02	27.03

Tabela B.16: MAPE obtido na fase de avaliação para os algoritmos de séries temporais com a variância, em cheias

Modelos	Horizonte									
	1	2	3	4	5	6	7	8	9	10
VAR	11.87	12.74	15.32	17.28	21.71	25.38	28.67	31.32	33.38	35.58
DYNLM	11.58	12.56	14.54	14.65	17.29	17.89	18.95	18.41	19.79	19.84
BVAR	10.61	11.38	16.18	19.15	22.28	25.69	29.17	32.72	37.08	40.21
RF	9.42	11.89	14.66	18.94	22.60	24.70	27.65	28.59	29.58	31.30
NN	41.02	28.68	30.80	32.38	31.04	23.52	27.74	30.73	33.24	38.61
CRF	9.72	9.28	11.42	11.18	15.09	15.56	18.60	17.01	18.17	18.37
naive	8.19	8.32	10.01	11.12	14.28	15.36	18.69	18.01	20.27	21.06

Tabela B.17: RMSE obtido na fase de avaliação para os algoritmos de séries temporais com granularidade mais fina

Modelos	Horizonte									
	1	2	3	4	5	6	7	8	9	10
VAR	18.51	26.23	32.86	37.89	40.85	42.38	42.84	42.63	41.97	41.11
DYNLM	19.13	21.13	23.31	25.26	26.80	27.96	28.89	29.59	30.11	30.42
BVAR	17.51	19.51	21.27	22.40	23.21	23.84	24.27	24.66	24.88	25.03
RF	23.59	25.87	27.83	29.19	30.54	30.60	30.08	31.77	32.48	32.68
NN	26.14	35.68	40.68	45.06	49.24	53.54	54.57	54.47	53.23	53.90
CRF	24.68	26.00	28.28	29.77	31.04	31.57	31.27	33.03	33.71	33.99
naive	21.05	22.67	24.35	26.05	26.99	27.85	28.28	28.99	29.45	29.62

Tabela B.18: RMSE obtido na fase de avaliação para os algoritmos de séries temporais com granularidade mais fina, em cheias

Modelos	Horizonte									
	1	2	3	4	5	6	7	8	9	10
VAR	16.17	23.13	27.00	26.21	28.48	31.91	33.64	34.36	35.78	35.77
DYNLM	21.30	24.81	26.99	26.18	27.78	30.13	31.42	30.56	31.06	29.71
BVAR	13.05	17.19	19.62	21.51	26.49	29.01	30.54	30.58	31.69	31.90
RF	20.01	29.21	34.44	37.60	40.75	42.49	44.50	45.39	46.86	48.43
NN	20.81	30.31	35.42	38.64	42.84	44.01	46.67	50.58	60.08	71.93
CRF	15.25	18.58	22.32	25.23	28.89	30.47	32.58	33.75	35.02	36.58
naive	13.64	14.13	16.10	17.00	20.43	21.14	25.03	23.73	26.02	27.03

Tabela B.19: MAPE obtido na fase de avaliação para os algoritmos de séries temporais com granularidade mais fina, em cheias

Modelos	Horizonte									
	1	2	3	4	5	6	7	8	9	10
VAR	9.15	12.46	14.67	15.41	17.87	19.90	21.45	21.94	23.39	24.40
DYNLM	11.37	13.85	15.22	16.18	18.21	19.74	21.56	21.09	22.02	21.13
BVAR	8.00	10.27	11.88	13.22	17.38	18.13	20.10	18.95	21.55	21.52
RF	13.83	21.50	25.96	28.85	32.47	34.34	37.31	38.76	40.28	42.08
NN	11.23	16.97	20.51	22.57	26.06	28.27	30.67	34.58	40.30	46.37
CRF	9.88	12.91	16.13	19.10	22.52	24.08	26.15	27.39	28.69	30.09
naive	8.19	8.32	10.01	11.12	14.28	15.36	18.69	18.01	20.27	21.06

Tabela B.20: RMSE obtido na fase de avaliação para os algoritmos de séries temporais com quantis

Modelos	Horizonte									
	1	2	3	4	5	6	7	8	9	10
DYNLM	20.29	23.18	26.64	29.92	32.23	34.77	36.98	38.83	40.72	42.37
RF	25.32	32.18	38.62	42.04	45.57	50.42	56.41	65.01	75.59	83.64
NN	21.30	23.80	26.26	27.41	28.69	29.57	29.88	30.79	31.02	31.12
naive	21.05	22.67	24.35	26.05	26.99	27.85	28.28	28.99	29.45	29.62

Tabela B.21: RMSE obtido na fase de avaliação para os algoritmos de séries temporais com quantis, em cheias

Modelos	Horizonte									
	1	2	3	4	5	6	7	8	9	10
DYNLM	28.01	32.89	41.13	51.61	59.61	67.90	75.51	80.93	87.22	92.74
RF	28.17	38.63	49.29	62.23	77.54	103.2	146.0	232.3	336.6	381.4
NN	28.53	28.38	30.69	31.11	34.40	36.60	39.25	39.87	41.24	42.09
naive	13.64	14.13	16.10	17.00	20.43	21.14	25.03	23.73	26.02	27.03

Tabela B.22: MAPE obtido na fase de avaliação para os algoritmos de séries temporais com quantis, em cheias

Modelos	Horizonte									
	1	2	3	4	5	6	7	8	9	10
DYNLM	16.15	19.21	24.36	31.68	37.58	43.60	50.74	55.12	60.21	64.73
RF	17.73	25.60	33.13	41.86	52.84	69.46	101.1	175.7	278.2	327.7
NN	14.27	14.45	17.36	17.89	22.15	23.64	26.75	27.38	28.77	29.36
naive	8.19	8.32	10.01	11.12	14.28	15.36	18.69	18.01	20.27	21.06

Tabela B.23: RMSE obtido na fase de avaliação para os algoritmos de séries temporais com pesos

Modelos	Horizonte									
	1	2	3	4	5	6	7	8	9	10
DYNLM	17.96	19.72	21.28	22.09	22.68	23.21	23.58	23.86	24.07	24.14
NN	20.92	24.88	28.55	32.87	31.09	28.86	27.92	28.29	28.67	28.08
CRF	19.59	24.90	28.18	30.89	35.30	33.13	36.15	38.13	37.26	38.46
naive	21.05	22.67	24.35	26.05	26.99	27.85	28.28	28.99	29.45	29.62

Tabela B.24: RMSE obtido na fase de avaliação para os algoritmos de séries temporais com pesos, em cheias

Modelos	Horizonte									
	1	2	3	4	5	6	7	8	9	10
DYNLM	22.33	25.87	28.38	28.02	30.19	31.84	33.00	32.23	32.99	32.48
NN	110.26	102.66	67.65	76.99	67.51	74.64	73.53	69.18	61.73	58.85
CRF	28.04	25.22	26.77	29.59	28.88	30.65	31.28	31.18	31.65	30.88
naive	13.64	14.13	16.10	17.00	20.43	21.14	25.03	23.73	26.02	27.03

Tabela B.25: MAPE obtido na fase de avaliação para os algoritmos de séries temporais com pesos, em cheias

Modelos	Horizonte									
	1	2	3	4	5	6	7	8	9	10
DYNLM	11.61	13.22	15.55	15.81	18.61	19.23	20.58	19.57	20.83	20.83
NN	53.84	46.27	37.00	47.09	45.19	50.20	51.06	46.65	42.36	41.08
CRF	18.72	15.98	17.63	20.46	19.86	21.17	21.79	23.81	22.57	23.56
naive	8.19	8.32	10.01	11.12	14.28	15.36	18.69	18.01	20.27	21.06

Tabela B.26: RMSE obtido em simulação

Modelos	Horizonte									
	1	2	3	4	5	6	7	8	9	10
DYNLM	17.99	18.27	18.78	19.30	19.62	19.87	20.12	20.30	20.49	20.54
RF	14.71	18.35	18.72	19.47	19.97	20.36	20.80	21.07	21.31	21.45
NN	15.28	19.19	19.77	21.08	22.73	23.54	24.48	24.20	24.16	24.38
CRF	15.59	19.50	20.06	20.66	21.13	21.41	21.88	22.34	22.45	22.68
naive	21.05	22.67	24.35	26.05	26.99	27.85	28.28	28.99	29.45	29.62

Tabela B.27: RMSE obtido em simulação, em cheias

Modelos	Horizonte									
	1	2	3	4	5	6	7	8	9	10
DYNLM	21.43	23.81	25.66	26.73	28.07	29.12	29.47	28.75	28.90	28.26
RF	13.32	14.95	16.94	16.70	19.27	21.22	23.31	22.95	23.83	24.76
NN	53.17	83.10	84.99	106.1	121.6	131.8	143.7	136.9	136.9	129.1
CRF	13.91	14.37	17.23	17.79	21.55	22.36	25.63	23.68	24.86	25.68
naive	13.64	14.13	16.10	17.00	20.43	21.14	25.03	23.73	26.02	27.03

Tabela B.28: MAPE obtido em simulação, em cheias

Modelos	Horizonte									
	1	2	3	4	5	6	7	8	9	10
DYNLM	11.72	13.41	15.31	16.11	17.56	17.96	18.65	17.54	17.62	17.33
RF	7.84	9.14	10.91	11.43	14.20	15.52	17.59	17.69	18.22	18.84
NN	19.74	25.20	32.23	39.77	50.12	56.23	62.87	62.39	66.50	63.23
CRF	8.20	8.90	11.16	11.81	15.45	15.36	18.66	17.57	18.30	18.88
naive	8.19	8.32	10.01	11.12	14.28	15.36	18.69	18.01	20.27	21.06

Tabela B.29: RMSE obtido em simulação com as variações

Modelos	Horizonte									
	1	2	3	4	5	6	7	8	9	10
DYNLM	20.81	20.37	21.23	21.75	21.73	22.14	22.33	22.60	22.96	22.95
RF	16.02	19.17	21.74	23.82	25.09	26.04	27.08	27.64	28.04	28.29
NN	18.93	23.50	26.70	30.25	33.37	36.43	39.47	42.81	46.03	49.14
CRF	19.29	22.39	24.90	28.03	30.63	33.24	35.78	38.59	41.35	43.96
naive	21.05	22.67	24.35	26.05	26.99	27.85	28.28	28.99	29.45	29.62

Tabela B.30: RMSE obtido em simulação com as variações em cheias

Modelos	Horizonte									
	1	2	3	4	5	6	7	8	9	10
DYNLM	15.09	14.35	15.66	17.65	19.33	19.83	23.06	23.04	24.58	25.27
RF	12.31	14.12	16.21	17.43	20.42	21.87	25.05	24.71	26.34	27.45
NN	15.80	23.30	32.23	35.49	39.49	44.70	50.95	54.83	59.66	62.98
CRF	13.05	12.78	16.01	18.05	22.73	24.38	30.19	30.89	34.36	36.63
naive	13.64	14.13	16.10	17.00	20.43	21.14	25.03	23.73	26.02	27.03

Tabela B.31: MAPE obtido em simulação com as variações, em cheias

Modelos	Horizonte									
	1	2	3	4	5	6	7	8	9	10
DYNLM	9.12	8.73	9.54	10.81	12.57	13.48	15.94	15.57	16.80	17.62
RF	6.48	8.21	10.14	11.56	14.55	14.88	18.42	18.39	19.70	20.67
NN	9.30	11.95	16.88	21.20	25.98	29.78	36.38	40.32	45.04	48.74
CRF	7.72	7.49	10.28	12.11	16.46	17.71	22.35	23.57	26.57	28.84
naive	8.19	8.32	10.01	11.12	14.28	15.36	18.69	18.01	20.27	21.06

Bibliografia

[1] C. Madureira and V. Baptista, *Hidroelectricidade em Portugal : memória e desafio*. Lisboa: Rede Eléctrica Nacional, S.A., 2002.

[2] Associação de Energias Renováveis, "A Eletricidade de Origem Renovável em Portugal Continental Novembro de 2014," Tech. Rep., 2014.

[3] Entidade Reguladora dos Serviços Energéticos, "Envolvente de Mercado," 2009. [Online]. Available: http://www.erse.pt/pt/supervisaodemercados/mercadodeelectricidade/envolventedemercado/Paginas/default.aspx?master=ErsePrint.master [Accessed: 2015-06-28]

[4] C. Ramos, M. Leal, and P. Silva, "Impactes das barragens nos regimes fluviais: comparação entre Vilarinho das Furnas (Hidroeléctrica) e Monte Novo (Hidroagrícola)," *Trunfos de uma Geografia Activa: desenvolvimento local, ambiente, ordenamento e tecnologia*, 2011.

[5] EDP Produção - Departamento de Gestão da Operação, "Alto Rabagão: Normas de Exploração da Albufeira e Normas de Descarregamento," EDP PRODUÇÃO, (Documento Confidencial), Tech. Rep.

[6] ——, "Pocinho: Normas de Exploração da Albufeira e Normas de Descarregamento," EDP PRODUÇÃO, (Documento Confidencial), Tech. Rep.

[7] ——, "Valeira: Normas de Exploração da Albufeira e Normas de Descarregamento," EDP PRODUÇÃO, (Documento Confidencial), Tech. Rep.

[8] ——, "Venda Nova: Normas de Exploração da Albufeira e Normas de Descarregamento," EDP PRODUÇÃO, (Documento Confidencial), Tech. Rep.

[9] ——, "Salamonde: Normas de Exploração da Albufeira e Normas de Descarregamento," EDP PRODUÇÃO, (Documento Confidencial), Tech. Rep.

[10] M. T. L. Barros, F. T.-C. Tsai, S.-l. Yang, J. E. G. Lopes, and W. W.-G. Yeh, "Optimization of Large-Scale Hydropower System Operations," *Journal of Water Resources Planning and Management*, vol. 129, no. 3, pp. 178–188, 2003.

[11] H. I. Jager and B. T. Smith, "Sustainable reservoir operation: can we generate hydropower and preserve ecosystem values?" *River research and Applications*, vol. 24, no. 3, pp. 340–352, 2008.

[12] S. Vedula and P. Mujumdar, "Optimal reservoir operation for irrigation of multiple crops," *Water Resources Research*, vol. 28, no. 1, pp. 1–9, 1992.

[13] StatSoft Inc, "What is Data Mining (Predictive Analytics, Big Data)," 2015. [Online]. Available: http://www.statsoft.com/Textbook/Data-Mining-Techniques#mining [Accessed: 2015-03-01]

[14] Portal Energia, "Funcionamento da energia hidrica e Barragens Hidroelectricas," 2008. [Online]. Available: http://www.portal-energia.com/funcionamento-da-energia-hidrica-barragens-hidroelectricas/ [Accessed: 2014-01-24]

[15] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, second edi ed. Elsevier, 2006.

[16] R. Roiger and M. Geatz, *Data Mining: A Tutorial-based Primer*. Addison Wesley, 2003.

[17] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, "CRISP-DM 1.0 Step-by-step data mining guide," 2000. [Online]. Available: http://the-modeling-agency.com/crisp-dm.pdf [Accessed: 2015-01-05]

[18] F. Bessler, D. Savic, and G. Walters, "Water reservoir control with data mining," *Journal of Water Resources Planning and Management*, vol. 129, no. 1, pp. 26–34, Jan. 2003. [Online]. Available: http://ascelibrary.org/doi/abs/10.1061/(ASCE)0733-9496(2003)129:1(26)

[19] M. I. Hejazi and X. Cai, "Building more realistic reservoir optimization models using data mining – A case study of Shelbyville Reservoir," *Advances in Water Resources*, vol. 34, no. 6, pp. 701–717, Jun. 2011. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0309170811000509

[20] A. Mohan and P. Z. Revesz, "Applications of spatio-temporal data mining to north platter river reservoirs," in *Proceedings of the 18th International Database Engineering & Applications Symposium*. New York, New York, USA: ACM Press, 2014, pp. 306–309. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2628194.2628221

[21] W. Ishak, "Modelling Reservoir Water Release Decision Using Temporal Data Mining and Neural Network," *International Journal of Emerging Technology and Advanced Engineering*, vol. 2, no. 8, pp. 422–428, 2012.

[22] A. Mohan and P. Revesz, "Temporal data mining of uncertain water reservoir data," in *Proceedings of the Third ACM SIGSPATIAL International Workshop on Querying and Mining Uncertain Spatio-Temporal Data*. New York, New York, USA: ACM Press, 2012, pp. 10–17. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2442985.2442987

[23] M. I. Hejazi, X. Cai, and B. L. Ruddell, "The role of hydrologic information in reservoir operation – Learning from historical releases," *Advances in Water Resources*, vol. 31, no. 12, pp. 1636–1650, Dec. 2008. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0309170808001309

[24] J. D.Cryer and K.-S. Chan, *Time Series Analysis With Applications in R*, second edi ed., G. Casella, S. Fienberg, and I. Olkin, Eds. Springer, 2008.

[25] P. J. Brockwell and R. A. Davis, *Introduction to Time Series and Forecasting*, second edi ed., G. Casella, S. Fienberg, and I. Olkin, Eds. Taylor & Francis, 2002.

[26] A. Coghlan, "Using R for Time Series Analysis," 2010. [Online]. Available: http://a-little-book-of-r-for-time-series.readthedocs.org/en/latest/src/timeseries.html [Accessed: 2015-01-09]

[27] R. H. Shumway and D. S. Stoffer, *Time Series Analysis and Its Applications*, third edit ed., G. Casella, S. Fienberg, and I. Olkin, Eds. Springer Science & Business Media, 2011, vol. 97. [Online]. Available: http://www.springerlink.com/index/10.1007/978-1-4419-7865-3

[28] J. H. Stock, "Forecasting economic time series." *A Companion to Theoretical Econometrics*, pp. 562–84, 1999.

[29] G. Reinsel, *Elements of Multivariate Time Series Analysis*. Springer New York, 2012. [Online]. Available: https://books.google.pt/books?id=To3kBwAAQBAJ

[30] R. Koenker and K. F. Hallock, "Quantile Regression," *Journal of Economic Perspectives*, vol. 15, no. 4, pp. 143–156, 2001.

[31] C. Stover and E. W. Weisstein, "Quantile." [Online]. Available: http://mathworld.wolfram.com/Quantile.html [Accessed: 2015-06-12]

[32] A. Cannon, *Essential Statistics*. Springer US, 2001, vol. 55, no. 1.

[33] NCS Pearson, "First Quartile." [Online]. Available: http://math.tutorvista.com/statistics/first-quartile.html [Accessed: 2015-06-12]

[34] S. Despa, "Quantile Regression," *Cornell University, Cornell Statistical Consulting. Stat-News*, no. 70, 2007.

[35] C. Perlich, S. Rosset, R. D. Lawrence, and B. Zadrozny, "High-quantile modeling for customer wallet estimation and other applications," in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM Press, 2007, pp. 977–985. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1281192.1281297

[36] L. Torgo, R. Ribeiro, B. Pfahringer, and P. Branco, "SMOTE for Regression," in *Progress in Artificial Intelligence*, L. Correia, L. P. Reis, and J. Cascalho, Eds. Springer Berlin Heidelberg, 2013, pp. 378–389.

[37] C. Elkan, "The foundations of cost-sensitive learning," in *International joint conference on artificial intelligence*. Citeseer, 2001, pp. 973–978. [Online]. Available: http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780195325195.001.0001/acprof-9780195325195

[38] L. Torgo and R. Ribeiro, "Predicting rare extreme values," in *Advances in Knowledge Discovery and Data Mining*. Springer, 2006, pp. 816–820.

[39] G. Xu, Y. Zong, and Z. Yang, *Applied Data Mining*. CRC Press, 2013.

[40] Yale University, "Scatterplot." [Online]. Available: http://www.stat.yale.edu/Courses/1997-98/101/scatter.htm [Accessed: 2015-06-17]

[41] R. Rodrigues, C. Brandão, and J. P. da Costa, "Hidrologia das cheias do Mondego de 26 e 27 de Janeiro de 2001," *Relatório do INAG*, 2001.

[42] F. Botelho and N. Ganho, "Dinâmica anticiclónica subjacente à seca de 2004 / 2005 em Portugal Continental," *VI Seminário Latino Americano de Geografia Física*, pp. 1–14, 2010.

[43] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2015. [Online]. Available: http://www.r-project.org/

[44] The Pennsylvania State University, "Vector Autoregressive models VAR(p) models," 2015. [Online]. Available: https://onlinecourses.science.psu.edu/stat510/node/79 [Accessed: 2015-06-03]

[45] ——, "Moving Average Models (MA models)," 2015. [Online]. Available: https://onlinecourses.science.psu.edu/stat510/node/48 [Accessed: 2015-06-03]

[46] K. L. Priddy and P. E. Keller, *Artificial Neural Networks: An Introduction*, ser. Tutorial Text Series. SPIE Press, 2005.

[47] C. Zhang and Y. Ma, *Ensemble Machine Learning: Methods and Applications*. Springer, 2012. [Online]. Available: https://books.google.pt/books?id=CjAs4stLXhAC

[48] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.

[49] R. S. Tsay, *MTS: All-Purpose Toolkit for Analyzing Multivariate Time Series (MTS) and Estimating Multivariate Volatility Models*, 2015. [Online]. Available: http://cran.r-project.org/package=MTS

[50] P. Brandt, *MSBVAR: Markov-Switching, Bayesian, Vector Autoregression Models*, 2015. [Online]. Available: http://cran.r-project.org/package=MSBVAR

[51] G. Grothendieck, *dyn: Time Series Regression*, 2012. [Online]. Available: http://cran.r-project.org/package=dyn

[52] W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S*, 4th ed. New York: Springer, 2002. [Online]. Available: http://www.stats.ox.ac.uk/pub/MASS4

[53] A. Liaw and M. Wiener, "Classification and Regression by randomForest," *R News*, vol. 2, no. 3, pp. 18–22, 2002. [Online]. Available: http://cran.r-project.org/doc/Rnews/

[54] T. Hothorn, P. Buehlmann, S. Dudoit, A. Molinaro, and M. V. D. Laan, "Survival Ensembles," *Biostatistics*, vol. 7, no. 3, pp. 355–373, 2006.

[55] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution," *BMC Bioinformatics*, vol. 8, no. 25, 2007. [Online]. Available: http://www.biomedcentral.com/1471-2105/8/25

[56] C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis, "Conditional Variable Importance for Random Forests," *BMC Bioinformatics*, vol. 9, no. 307, 2008. [Online]. Available: http://www.biomedcentral.com/1471-2105/9/307

[57] J. Gattorna, *Strategic Supply Chain Alignment: Best Practice in Supply Chain Management*. Gower, 1998.

[58] B. A. Olshausen, "Aliasing," Tech. Rep., 2000.

[59] MathWorks, "var." [Online]. Available: http://www.mathworks.com/help/matlab/ref/var.html [Accessed: 2015-06-10]

[60] H. Yu and N. A. A. Rahim, *Imaging in Cellular and Tissue Engineering*, ser. Series in Cellular and Clinical Imaging. Taylor & Francis, 2013.

[61] D. G. Pascual, *Artificial Intelligence Tools: Decision Support Systems in Condition Monitoring and DIagnosis*. CRC Press, 2015.

[62] K. Diamantaras, W. Duch, and L. S. Iliadis, *Artificial Neural Networks - ICANN 2010: 20th International Conference, Thessaloniki, Greece, September 15-18, 2010, Proceedings*, ser. Artificial Neural Networks - ICANN 2010: 20th International Conference, Thessaloniki, Greece, September 15-18, 2010 : Proceedings. Springer, 2010.

[63] M. Smithson and E. C. Merkle, *Generalized Linear Models for Categorical and Continuous Limited Dependent Variables*, ser. Chapman & Hall/CRC Statistics in the Social and Behavioral Sciences. Taylor & Francis, 2013.

[64] R. Koenker, *quantreg: Quantile Regression*, 2015. [Online]. Available: http://cran.r-project.org/package=quantreg

[65] A. J. Cannon, “Quantile regression neural networks: implementation in R and application to precipitation downscaling,” *Computers & Geosciences*, vol. 37, pp. 1277–1284. doi:10.1016/j.cageo.2010.07.005, 2011.

[66] N. Meinshausen, *quantregForest: Quantile Regression Forests*, 2012. [Online]. Available: http://cran.r-project.org/package=quantregForest